Numerical Integration of Differential-Algebraic Equations with Harmless Critical Points

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Mathematik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät II
Humboldt-Universität zu Berlin

von
M. Sc. Rakporn Dokchan

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II:
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Dr. R. März
2. Prof. Dr. W. Römisch
3. Prof. Dr. E. Weinmüller

eingereicht am: 12. Januar 2011
Tag der mündlichen Prüfung: 21. Mai 2011

Acknowledgement

My first and sincere thanks are given to my supervisor Prof. Dr. Roswitha März, for accepting me, giving me the chance to enter the world of differential-algebraic equations and to write this thesis, and also for her invaluable support and guidance.

I especially want to thank Asst. Prof. Dr. Pichan Sawangwong for suggesting me to study in Berlin and Dr. Rene Lamour for fruitful discussions and suggestions on implementation issues.

Special thanks go to Stefan Vigerske, my office mate at the Humboldt University of Berlin, for helpful review of the thesis and comments. I wish to thank all my colleagues at the Humboldt University of Berlin for the pleasant working atmosphere. I greatly appreciate the help of Patcharee Larpsuriyakul, Decha Dechtrirat, Jarungjit Parnjai, Maneenate Wechakama and Thitinan Tantidham.

As a scholar, I am very grateful to The Royal Thai Government for granting me the financial support to do the Ph.D. in Mathematics in Germany.

Eventually, I personally would like to express my gratitude to my parents, my brothers and sisters for their warm understanding, encouraging, and assistance during my long stay in Germany. Most importantly, I wish to thank my wonderful sister, Rasri Thovasakul (Dokchan), for her never-ending family support.

Rakporn Dokchan

Berlin, January 2011

Abstract

Differential-algebraic equations (DAEs) are implicit singular ordinary differential equations, which describe dynamical processes that are restricted by some constraints. In contrast to explicit regular ordinary differential equations, for a DAE not any value can be imposed as an initial condition. The initial values need to be consistent with the DAE. Furthermore, DAEs involve not only integration problems but also differentiation problems. The differentiation index of a DAE indicates the number of differentiations required in order to solve a DAE.

Since approximately 1980, DAEs form a research area of applied mathematics, which primarily focuses on the characterization and classification of regular problem classes and the construction and foundation of integration methods for simulation software. Among others, S.L. Campbell, L.R. Petzold, E. Griepentrog, R. März, W.R. Rheinboldt, P. Rabier, E. Hairer, Ch. Lubich, V. Mehrmann, P. Kunkel, und R. Riaza have made significant contributions for this purpose.

The numerical treatment of DAEs requires knowledge about their structure. I. Higueras, R. März, and C. Tischendorf have shown in 2003 that one can reliably integrate a general linear DAE with a properly stated leading term,

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \qquad t \in \mathcal{J}, \tag{1}$$

which is regular with tractability index 2 – in contrast to linear DAEs in standard form.

The first study of the classification of critical points of linear DAEs have been published by R. Riaza and R. März in 2004-2008. Based on the tractability index, critical points are classified according to failures of certain rank conditions of matrix functions. Essentially, a critical point is said to be harmless, if the flow described by the inherent differential equation is not affected.

The subject of this work are linear DAEs of the form (1). Index-2 DAEs with harmless critical points are characterized. Under the application of quasi-admissible projector functions instead of the admissible ones, besides DAEs which have almost everywhere the same characteristic values, DAEs with index changes can now be discussed for the first time. The main part of the work is to provide a proof of feasibility, convergence, and only weak instability of numerical integration methods (BDF, IRK (DAE)) for general linear index-2 DAEs with harmless critical points, as well as the development and testing of error estimators and stepsize control.

Zusammenfassung

Algebro-Differentialgleichungen (engl. differential-algebraic equations – DAEs) sind implizite singuläre gewöhnliche Differentialgleichungen, die dynamische Prozesse, die Restriktionen unterliegen, beschreiben. Sie unterscheiden sich von expliziten gewöhnlichen Differentialgleichungen dahingehend, dass Anfangswerte nicht beliebig vorgegeben werden können. Sie müssen konsistent mit der DAE sein. Darüberhinaus sind in einer DAE sowohl Integrations- als auch Differentiationsaufgaben involviert. Der Differentiationsindex einer DAE gibt an, wieviele Differentiationen zur Lösung der DAE notwendig sind.

DAEs bilden seit etwa 1980 ein Arbeitsgebiet der Angewandten Mathematik, wobei es vorwiegend um die Charakterisierung und Klassifizierung regulärer Aufgabenklassen und die Konstruktion nebst Fundierung von Integrationsmethoden für Simulationssoftware geht. Unter anderen haben S.L.Campbell, L.R.Petzold, E.Griepentrog, R.März, W.R.Rheinboldt, P.Rabier, E.Hairer, Ch.Lubich, P.Kunkel, V.Mehrmann, und R.Riaza hierzu wichtige Beiträge geleistet.

Die numerische Behandlung von DAEs erfordert Kenntnisse über deren Struktur. I.Higueras, R.März und C.Tischendorf haben 2003 gezeigt, dass man allgemeine lineare DAEs mit properem Hauptterm,

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \qquad t \in \mathcal{J}, \tag{2}$$

die regulär mit Traktabilitätsindex 2 sind, zuverlässig numerisch integrieren kann – im Unterschied zu linearen DAEs in Standardform.

Erste Arbeiten zur Klassifizierung von kritischen Punkten bei linearen DAEs wurden von R.Riaza und R.März 2004-2008 publiziert. Formal werden kritische Punkte an die Verletzung bestimmter Rangbedingungen von Matrixfunktionen im Rahmen des Traktabilitätsindexes geknüpft. Im wesentlichen heißt ein kritischer Punkt harmlos, wenn der durch die inhärente Differentialgleichung beschriebene Fluß nicht tangiert ist.

Gegenstand der vorliegenden Arbeit sind lineare DAEs der Form (2). Es werden Index 2 DAEs mit harmlosen kritischen Punkten charakterisiert. Unter Verwendung von quasi-zulässigen Projektorfunktionen statt der zulässigen können neben DAEs, die fast überall gleiche charakteristische Werte haben, nun erstmalig auch solche mit Indexwechseln behandelt werden. Der Hauptteil der Arbeit besteht im Nachweis von Durchführbarkeit, Konvergenz und nur schwacher, strukturell beschränkter Instabilität von numerischen Integrationsmethoden (BDF, IRK(DAE)) für allgemeine lineare Index 2 DAEs mit harmlosen kritischen Punkten, sowie in der Entwicklung und Erprobung von Fehlerschätzer und Schrittweitensteuerung.

Contents

List of Figures

Chapter 1

Introduction

An implicit ordinary differential equation (ODE) has the form

$$f(x'(t), x(t), t) \;=\; 0, \tag{1.1}$$

with $f : \mathbb{R}^m \times \mathcal{D}_f \times \mathcal{I}_f \to \mathbb{R}^m$, where $x : \mathcal{I}_f \to \mathbb{R}^m$ denotes the unknown function, $\mathcal{D}_f \subseteq \mathbb{R}^m$ is an open subset, and $\mathcal{I}_f \subseteq \mathbb{R}$ is an interval. If the partial derivative $f'_y(y, x, t)$ is nonsingular for all values of its arguments, then equation (1.1) is locally equivalent to an explicit ODE $x' = \varphi(x, t)$. An implicit ODE (1.1) with $f'_y(y, x, t)$ everywhere singular is called a differential-algebraic equation (DAE).

DAEs describe dynamical processes that are restricted by some constraints. DAEs are distinguished from explicit regular ODEs in several aspects. An important characteristic of DAEs is that not any value can be imposed as an initial condition (cf. Example 2.1). Some components of the solution are determined by the algebraic equations or constraints. Furthermore, the dynamics of the problem is in fact determined by a lower dimensional ODE, sometimes called underlying ODE or inherent regular ODE.

Since approximately 1980, DAEs form a research area of applied mathematics, which primarily focuses on the characterization and classification of regular problem classes and the construction and foundation of integration methods for simulation software. Among others, S.L. Campbell, L.R. Petzold, E. Griepentrog, R. März, W.R. Rheinboldt, P. Rabier, E. Hairer, Ch. Lubich, V. Mehrmann, P. Kunkel, und R. Riaza have made significant contributions for this purpose [9, 27, 35, 49, 75, 76].

In this work, we consider a linear time-varying DAE of the form

$$A(t)(D(t)x(t))' + B(t)x(t) \;=\; q(t), \qquad t \in \mathcal{J}. \tag{1.2}$$

The matrix functions A, D, B, and the right-hand side q are assumed to be continuous on the interval $\mathcal{J} \subseteq \mathbb{R}$. Thereby, the coefficient D has constant rank on $\mathcal{J}$ and describes the derivative component of the unknown function. Under certain conditions indicated later, equation (1.2) will be said to be a DAE with a *properly stated leading term* (cf. [2, 41, 63]) or a DAE with a *quasi-proper leading term* (cf. [54, 67]).

A standard linear time-varying DAE

$$E(t)x'(t) + F(t)x(t) \;=\; q(t), \tag{1.3}$$

with E, $F \in L(\mathbb{R}^m)$ can be rewritten, as proposed in [54], in the form (1.2). For instance, if there exists a continuously differentiable projector function P_E such that $\ker P_E \subseteq \ker E$, then, since $E = EP_E$, we can reformulate (1.3) as

$$E(t)(P_E(t)x(t))' + (F(t) - E(t)P_E'(t))x(t) \;=\; q(t), \tag{1.4}$$

which takes the form (1.2) with $D(t) = P_E(t)$, $B(t) = F(t) - E(t)P_E'(t)$. Such a projector P_E exists if the matrix E is in $C^1(\mathcal{I}, L(\mathbb{R}^m))$ and has constant rank [2, 41]. A projector P_E may also exist in the case when the rank of E varies [54].

The definition of the tractability index of a DAE (1.2) is based on the construction of a matrix chain via a sequence of suitably chosen projector functions in such a way that a decoupling of the dynamic and (possibly hidden) algebraic components is achieved, see Chapter 2 for an exact definition. The DAE (1.2) is said to be *regular* if the involved matrices, projector functions, and associated subspaces satisfy certain criteria like constant rank, differentiability, and transversality, respectively, see Definition 2.15. The regularity of the DAE ensures that consistent initial values allow for a smooth flow of the dynamical components of the DAE's solution. A point t_* is said to be regular if there exists an open interval that contains t_* and where the DAE is regular.

Roughly speaking, a *critical* point of a DAE is characterized by a failure of one of the regularity criteria. Several unusual phenomena may occur in context of critical points, e.g., one may associate with critical points the non-existence or the non-uniqueness of the solutions to the DAE system.

Example 1.1. Consider the DAE

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \Big(\begin{bmatrix} 1 & -t \end{bmatrix} x(t) \Big)' + \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} x(t) \;=\; 0, \quad t \in \mathcal{J} = \mathbb{R}, \tag{1.5}$$

which has the form (1.2) with $A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $D = \begin{bmatrix} 1 & -t \end{bmatrix}$, $B = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, $m = 2$, $n = 1$.
On the intervals $(-\infty, 1)$ and $(1, \infty)$ the DAE solutions are given by

$$x(t) \;=\; \frac{1}{1-t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t),$$

where u satisfies the inherent singular ODE

$$u'(t) + \frac{2}{1-t} u(t) \;=\; 0. \tag{1.6}$$

We consider $t_* = 1$ to be a critical point. The homogeneous ODE (1.6) has the solutions

$u(t) = (t-1)^2 u(0)$. Then, the solutions of the equation (1.5) are

$$x(t) \;=\; (1-t)u(0)\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad u(0) \in \mathbb{R},$$

which show that all solutions vanish at $t_* = 1$. Uniqueness of solutions is therefore lost at the critical point.

In a projector-based framework, a point where the matrix functions $A(t)D(t)$ or $E(t)$ change their rank on the considered interval is also defined as a critical point.

Example 1.2. The DAE

$$\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} x'(t) + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x(t) \;=\; q(t) \tag{1.7}$$

with scalar continuous function α on the interval $\mathcal{I} = [-1, 1]$,

$$\alpha(t) = 0 \quad \text{for} \quad t \le 0, \quad \alpha(t) \neq 0 \quad \text{for} \quad t > 0,$$

has exactly one solution

$$\begin{aligned} x_2(t) \;&=\; q_2(t), \quad t \in [-1, 1] \\ x_1(t) \;&=\; \begin{cases} q_1(t), & t \in [-1, 0], \\ q_1(t) - \alpha(t) q_2'(t), & t \in (0, 1]. \end{cases} \end{aligned}$$

The point $t_* = 0$ where the matrix $E(t)$ changes its rank is a critical one. The solvability statements for regular DAEs show that functions q_2 are continuous on the entire interval $\mathcal{I}$ and continuously differentiable on $(0, 1]$. For instance, if

$$\alpha(t) = \begin{cases} 0\ , & t \in [-1, 0], \\ t^{\frac{1}{3}}, & t \in (0, 1], \end{cases}$$

then for $q_1(t) = 0$ and $q_2(t) = \alpha(t)$ we obtain

$$x_2(t) = \alpha(t) \quad \text{and} \quad x_1(t) = \begin{cases} 0, & t \in [-1, 0], \\ -\frac{1}{3} t^{-\frac{1}{3}}, & t \in (0, 1]. \end{cases}$$

As shown in Figure 1.1, the solution segments of the second component can be glued

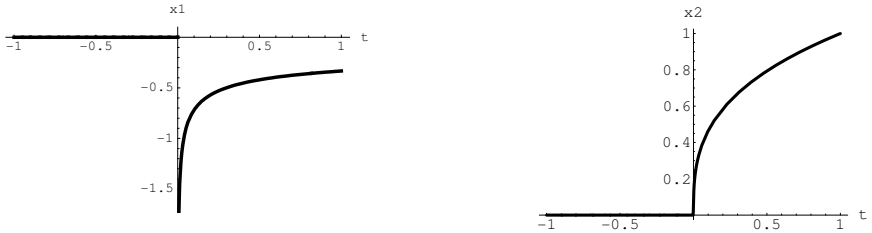


Figure 1.1: The solutions x_1, x_2 in case of $q_1 = 0$ and $q_2 = \alpha$

together smoothly, whereas this is not possible for the first component. If we relax the strong solvability concept and choose smoother functions q, a continuous solution may be available on the whole interval $\mathcal{I}$. For example, for $q_1(t) = 0$, $q_2(t) = t^2$, the particular solution is $x_2(t) = t^2$ and $x_1(t) = \begin{cases} 0\,, & t \in [-1, 0], \\ -2t^{\frac{4}{3}}, & t \in (0, 1]. \end{cases}$ See Figure 1.2.

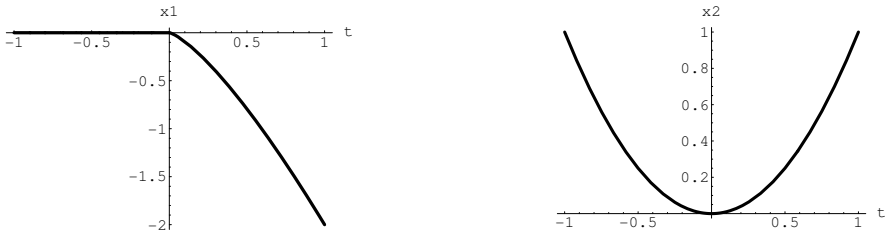


Figure 1.2: The solutions x_1, x_2 in case of smoother function q

More precisely, if we rewrite the DAE (1.7) as

$$\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} x \right)' + x \;=\; q, \tag{1.8}$$

we see that continuous differentiability is required only for the component x_2, which can be ensured by assuming that q_2 is continuously differentiable on $[-1, 1]$.

Observe that the critical point of the DAE (1.5) yields a singularity within the inherent ODE. In order to solve this kind of equation a special treatment is necessary. In [47] the convergence behavior of collocation schemes applied to approximate solutions of boundary value problems (BVPs) in linear index-1 DAEs which exhibit such a critical point of the DAE has been studied. Our attention is, however, directed to a critical point which can be heeled by higher smoothness as considered in Example 1.2. Such a critical point, being *harmless*, can be characterized by continuous extensions of certain projector functions or by quasi-regular DAEs.

The first study of the classification of critical points of linear DAEs have been published by R. Riaza and R. März in 2004-2008 [68, 69, 71, 77]. Based on the projector method, critical points have been classified according to failures of certain rank conditions of matrix functions [69]. These categories of critical points have been proven to be independent of the choice of (admissible) projector functions and to be invariant under linear time-varying coordinate changes and refactorizations. As stated in [69, 76], assuming the existence of continuous extensions of certain projector functions and density of the set of regular points, we can construct a critical matrix sequence and use it to DAE with critical points and characterize harmless critical points. As a consequence, if a DAE possesses only harmless critical points, singularity of the so-called inherent explicit regular ODE may be avoided [69]. The flow described by the inherent differential equation is not affected.

Unfortunately, the working assumptions proposed in [69, 76] present some limitations. If the tractability indices of DAEs are not uniform on the whole regular

intervals, one cannot define such harmless critical points. Nevertheless, for quasi-regular DAEs as proposed by März in [67] we may also characterize harmless critical points if the indices are not uniform, see Chapter 4. Under the application of quasi-regular projector functions instead of the regular ones, besides DAEs which have almost everywhere the same characteristic values, DAEs with index changes can now be discussed for the first time. The idea of these quasi-regular DAEs is to use the continuous subnullspace instead of the discontinuous nullspace of the matrix function for the construction of the matrix chain. Concerning again equation (1.3), if the coefficient E does not have a constant rank, we cannot define a C^1 projector function P_E along the kernel of E and hence, we may not rewrite (1.3) in the form (1.2) by means of the nullspace projector. However, if we choose the projector function $P_E \in C^1(\mathcal{I}, L(\mathbb{R}^m))$ with $\ker P_E \subseteq \ker E$ we may transform (1.3) to (1.2) with a quasi-proper leading term. This is a particular instance of a quasi-regular DAE.

The main purpose of this work is to study numerical integration methods for linear time-varying index-2 DAEs of the form (1.2) possessing harmless critical points. Numerical solutions of DAEs pose difficulties for standard ODE methods. Hence, numerous numerical integration methods developed for ODEs have been modified to approximate solutions of DAEs (see, e.g., [10], [34], [41]). Numerous convergence results were given in the literature. For index-1 DAEs in standard form (1.3), convergence results of the backward differentiation formulas (BDF) and Runge-Kutta methods have been presented, e.g., in [9, 27]. For index-2 DAEs, stability and convergence results for the BDF and Runge-Kutta schemes applied to the Hessenberg form have been described in [34]. For more general systems, the stability behavior of the solutions of linear-implicit index-2 DAEs

$$A(t)x'(t) + g(x,t) \;=\; 0$$

have been investigated in [89]. Thereby, the nullspace of the leading coefficient matrix $A(t)$ is assumed to be constant. Numerical integration applied to DAEs with properly stated leading terms has been studied in [40, 41, 43].

A well-known DAE solver code based on the BDF method is DASSL [72]. This code has been written for solving initial value problems in the standard form DAE (1.1) having index ≤ 1. RADAU5 is a DAE solver code developed in [32]. This code is based on the Radau IIA methods with stage number $s = 3$ [33] and is implemented for solving ODEs and semi-implicit index-1 DAEs.

I. Higueras, R. März, and C. Tischendorf have shown in 2003 [43] that one can reliably integrate a general linear DAE with a properly stated leading term (1.2) which is regular with tractability index 2 – in contrast to linear DAEs in standard form (1.3). The following example shows that a DAE given in the standard formulation (1.3) can cause serious difficulties. An appropriate formulation of the problem, for instance, in form of a properly stated DAE, is more preferable and ensures a correct behavior of the numerical solution when a standard ODE method

is applied to DAEs [2, 41, 42, 43].

Example 1.3. Consider the DAE in standard formulation

$$\begin{bmatrix} 0 & 0 \\ 1 & \zeta t \end{bmatrix} x'(t) + \begin{bmatrix} 1 & \zeta t \\ 0 & 1+\zeta \end{bmatrix} x(t) = \begin{bmatrix} g(t) \\ 0 \end{bmatrix}, \tag{1.9}$$

which is equivalent to the system

$$x_1(t) + \zeta t x_2(t) = g(t), \tag{1.10a}$$
$$x_1'(t) + \zeta t x_2'(t) + (1+\zeta) x_2(t) = 0, \tag{1.10b}$$

where ζ is a real number and $g(t)$ is a smooth function. The exact solution of this system is given by

$$x_1(t) = -\zeta t x_2(t) + g(t), \quad x_2(t) = -g'(t).$$

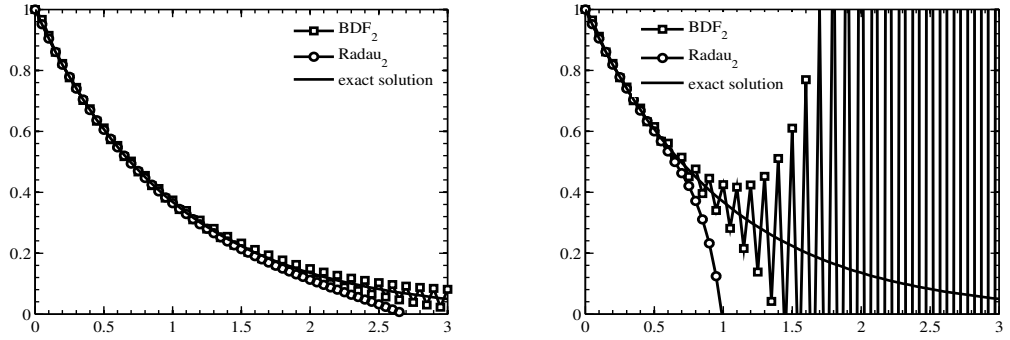
It is known [24, 25] that the implicit Euler method applied to this index-2 DAE fails completely for $\zeta = -1$. If $\zeta \neq -1$, the method is feasible but it converges only if $\left|\frac{\zeta}{1+\zeta}\right| < 1$, that is, if $\zeta > -0.5$. The implicit Euler method discretization of the DAE (1.10) with constant stepsize h is

$$x_{1,\ell} + \zeta t_\ell x_{2,\ell} = g(t_\ell), \tag{1.11a}$$
$$\frac{1}{h}\left(x_{1,\ell} - x_{1,\ell-1}\right) + \zeta t_\ell \frac{1}{h}\left(x_{2,\ell} - x_{2,\ell-1}\right) + (1+\zeta) x_{2,\ell} = 0. \tag{1.11b}$$

Inserting (1.11a) into (1.11b) yields the recursion

$$x_{2,\ell} = \frac{\zeta}{1+\zeta} x_{2,\ell-1} - \frac{1}{(1+\zeta)h}\left(g_\ell - g_{\ell-1}\right), \qquad \zeta \neq -1,$$

for the second component of the solution. The numerical solution x_2 for $g(t) = e^{-t}$ and $h = 0.05$ is shown in Figure 1.3. For different values of ζ all integration methods fail or provide unstable numerical solutions.



(a): DAE (1.10), $\zeta = -0.26$ (b): DAE (1.10), $\zeta = -0.3$

Figure 1.3: Numerical solution x_2 of DAE (1.10) for $g(t) = e^{-t}$, $h = 0.05$, and different values of ζ. $x^0 = (1,1)^T$ was used as a consistent initial value.

However, if we reformulate the DAE (1.10) into a DAE with a properly stated leading

term,

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 & \zeta t \end{bmatrix} x(t) \right)' + \begin{bmatrix} 1 & \zeta t \\ 0 & 1 \end{bmatrix} x(t) \;=\; \begin{bmatrix} g(t) \\ 0 \end{bmatrix}, \tag{1.12}$$

and apply the implicit Euler method to this formulation, we obtain

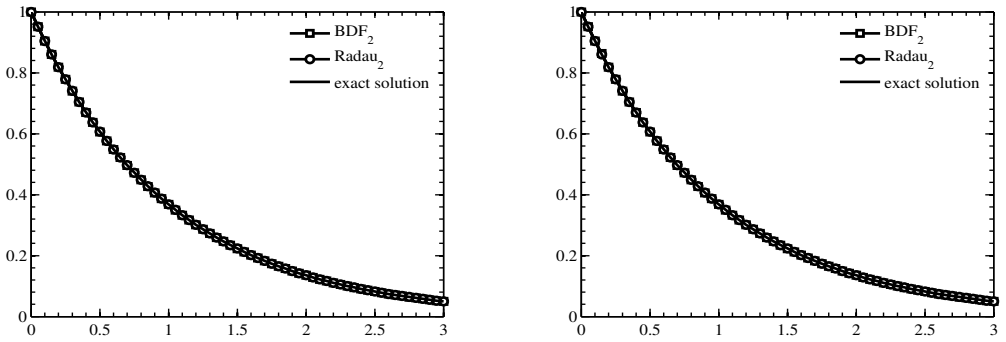
$$x_{1,\ell} + \zeta t_\ell x_{2,\ell} \;=\; g(t_\ell), \tag{1.13a}$$

$$\frac{1}{h}\left((x_{1,\ell} + \zeta t_\ell x_{2,\ell}) - (x_{1,\ell-1} + \zeta t_{\ell-1} x_{2,\ell-1})\right) + x_{2,\ell} \;=\; 0. \tag{1.13b}$$

Substituting (1.13a) into (1.13b) we obtain another recursion for x_2,

$$x_{2,\ell} \;=\; -\frac{1}{h}\left(g_\ell - g_{\ell-1}\right),$$

which shows that the implicit Euler method integrates correctly the exact solution $x_2(t) = -g'(t)$. In Figure 1.4 we give the numerical solution x_2 for the same values of ζ. In this case the numerical integration methods work well. It is important to note that the correct numerical results can be ensured by the properly stated leading term formulation.



(a): DAE (1.12), $\zeta = -0.26$ (b): DAE (1.12), $\zeta = -0.3$

Figure 1.4: Numerical solution x_2 of DAE (1.12) for $g(t) = e^{-t}$, $h = 0.05$, and different values of ζ. $x^0 = (1, 1)^T$ was used as a consistent initial value.

The main part of the work is to provide a proof of feasibility, convergence, and only weak instability of numerical integration methods (BDF, implicit Runge-Kutta method) for general linear index-2 DAEs with harmless critical points, as well as the development and testing of error estimators and stepsize control.

This thesis is organized as follows.

In Chapter 2 we introduce the concept of DAEs with properly stated leading term on a given interval and some additional basic notions. We define the regularity and describe the decoupling procedure for linear DAEs (1.2) with properly stated leading term. This decoupling procedure is the key tool for studying numerical integration methods.

In Chapter 3 critical points of the linear DAEs (1.2) are classified in context of

constant rank and transversality conditions which are not satisfied. We characterize a harmless critical point of a linear DAE (1.2) by assuming the existence of continuous extensions of certain projector functions from the regularity set to the entire interval.

The definition of DAEs with quasi-proper leading term and quasi-regularity for linear DAEs (1.2) is given in Chapter 4. For quasi-regular linear DAEs, we also define the notion of harmless critical points.

As we are interested in the numerical solution of initial value problems of linear index-2 DAEs with harmless critical points, we characterize in Chapter 5 index-2 DAEs with harmless critical points and address decoupling procedures in more detail.

Runge-Kutta and BDF schemes are applied to index-2 DAEs (1.2) with harmless critical points in Chapter 6. Using the decoupling procedure from Chapter 5 we investigate the stability and convergence properties of Runge-Kutta and BDF methods for those problems.

Finally, a local error estimation and stepsize prediction algorithm for BDF methods applied to linear index-2 DAE (1.2) is described in Chapter 7. Numerical results indicate very good performance of the proposed error estimate.

Chapter 2

Regular Differential-Algebraic Equations

Differential-algebraic equations are special implicit ordinary differential equations of the form

$$f(x'(t), x(t), t) = 0, \tag{2.1}$$

with $f : \mathbb{R}^m \times \mathcal{D}_f \times \mathcal{I}_f \to \mathbb{R}^m$, where $x : \mathcal{I}_f \to \mathbb{R}^m$ denotes the unknown function, $\mathcal{D}_f \subseteq \mathbb{R}^m$ is an open subset, and $\mathcal{I}_f \subseteq \mathbb{R}$ is an interval. Thereby, the partial derivative $f'_y(y, x, t)$ is singular for all values of its arguments. As stated in the introduction, if $f'_y(y, x, t)$ is nonsingular, then equation (2.1) is locally equivalent to an explicit ODE $x' = \varphi(x, t)$.

DAEs are distinct from explicit regular ODEs in many aspects, for instance, the initial values must be chosen to satisfy the algebraic equations of the system. Further, DAEs involve not only integration problems but also differentiation problems, as illustrated in the following examples.

Example 2.1. Consider the DAE

$$x_1(t) - x_2(t) + x_3(t) = 0, \tag{2.2a}$$
$$x'_2(t) - x_1(t) + x_2(t) = 0, \tag{2.2b}$$
$$x_1(t) = r(t). \tag{2.2c}$$

This system has a solution

$$\begin{aligned} x_1(t) &= r(t), \\ x_3(t) &= x_2(t) - r(t), \end{aligned}$$

if and only if x_2 is a solution of the explicit ODE

$$x'_2(t) = -x_2(t) + r(t). \tag{2.3}$$

Clearly, only certain components of the solution need to be differentiable. Here, only the smoothness of x_2 is required. For x_1 and x_3 it is sufficient to require continuity. Furthermore, when solving (2.2) as an initial value problem (IVP), only an initial condition for x_2 can be imposed. The remaining components x_1 and x_3 are determined by x_2 and the right-hand side $r(t)$. Initial conditions for these components need to be *consistent*, i.e., $x_1(t_0) = r(t_0)$ and $x_3(t_0) = x_2(t_0) - x_1(t_0)$.

Example 2.2. Consider the DAE

$$x_1'(t) + x_1(t) - x_2(t) = 0, \tag{2.4a}$$
$$x_1(t) = r(t). \tag{2.4b}$$

The solution of this system is given by

$$x_1(t) = r(t), \quad x_2(t) = r'(t) + r(t),$$

under the assumption that r is a differentiable function. Obviously, the solvability of this system relies upon the derivative of the right-hand side r and no initial condition is needed. In contrast to the problem in Example 2.1 where we have to *integrate* the ODE (2.3), here we need to *differentiate* x_1 in order to solve this system for x_2. That means, the solution process for this problem involves a differentiation (not an integration).

From Example 2.1 and 2.2 we may distinguish between DAEs and ODEs as follows:

(1) Some components of the solution are determined by the algebraic equations of the DAEs. This implies that the initial values need to be consistent with the DAEs, when solving the IVPs.

(2) Some parts of the right-hand side need to be differentiated in order to obtain a solution of the DAEs. That means DAEs involve not only integration problems but also differentiation problems.

In the next section we introduce the notations, definitions and some properties of projectors and subspaces that will be useful throughout this thesis. Since linear constant coefficient DAEs provide fundamental results to reveal the inner structure of the DAEs and to develop various index concepts for linear DAEs with time-varying coefficients, we will address their regularity in Section 2.2. Section 2.3 addresses the basic notations and definitions for the regularity of linear DAEs with properly stated leading term. This will help us to characterize critical points in Chapter 3.

2.1 Preliminary material

In this section we present the notations and some basic properties of the projectors and subspaces. Although we sometimes provide proofs, we make no attempt at completeness. For more details we refer to [4, 27, 92].

Projectors and basic subspaces

Definition 2.3. *A square matrix $Q \in L(\mathbb{R}^m)$ is called a* projector *if the relation $Q^2 = Q$ is satisfied. The projector Q will be called a projector* onto *a subspace $N \subseteq \mathbb{R}^m$ if $\operatorname{im} Q = N$ and it will be called a projector* along *a subspace $N \subseteq \mathbb{R}^m$ if $\ker Q = N$.*

If Q is a projector onto a subspace N, then $P := I - Q$ is a projection along N. In addition, the properties $P + Q = I$ and $QP = PQ = 0$ hold.

The following lemma shows a relation between the Kronecker index and the tractability index for regular DAEs with constant coefficients which will be addressed in the next section.

Lemma 2.4. *For given E, $F \in L(\mathbb{R}^m)$, a projector $Q_E \in L(\mathbb{R}^m)$ onto the subspace $N_E := \ker E$, and the subspace $S_{EF} := \{z \in \mathbb{R}^m : Fz \in \operatorname{im} E\}$, the following conditions are equivalent:*

(i) $N_E \cap S_{EF} = \{0\}$.

(ii) $N_E \oplus S_{EF} = \mathbb{R}^m$.

(iii) the matrix $E_1 := E + FQ_E$ is nonsingular.

(iv) $\{E, F\}$ form a regular matrix pencil with Kronecker index 1.

Further, if the matrix E_1 is nonsingular, then the projector Q_E^c onto N_E along S_{EF} has the form

$$Q_E^c \;:=\; Q_E E_1^{-1} F.$$

Q_E^c is said to be the canonical *projector onto N_E along S_{EF}.*

The proof for this statement can be found in Theorem A.13 and Lemma A.14 of [27].

Definition 2.5. *For $i \in \mathbb{N} \cup \{0\}$, a time-dependent subspace $L(t) \subseteq \mathbb{R}^l$, $t \in \mathcal{J}$, is said to be a C^i-subspace on $\mathcal{J}$ if $L(t)$ has constant dimension and is spanned by basis functions that belong to $C^i(\mathcal{J}, \mathbb{R}^l)$.*

The existence of a C^1-subspace of $\mathbb{R}^m$ with constant dimension defined on an interval $\mathcal{J} \subseteq \mathbb{R}$ is equivalent to the existence of a C^1 projector function onto (or along) this space [28].

Theorem 2.6. *A time-dependent subspace $N(t)$, $t \in \mathcal{J} \subseteq \mathbb{R}$, is a C^1-subspace if and only if there is a continuously differentiable projector $Q(t)$ onto $N(t)$.*

Proof : Let $N(t)$ be a C^1-subspace, i.e., there exist a basis $\{d_1(t), d_2(t), \ldots, d_k(t)\}$ such that

$$N(t) = \operatorname{span}\{d_1(t), d_2(t), \ldots, d_k(t)\}, \quad t \in \mathcal{J},$$

with $d_i \in C^1(\mathcal{J}, \mathbb{R}^m)$, $i = 1, \ldots, k$. Denoting by $\mathcal{H}(t)$ the $k \times k$ matrix consisting of the columns $d_1(t)$, $d_2(t)$, …, $d_k(t)$ we may construct a projection

$$Q(t) := \mathcal{H}(t)(\mathcal{H}(t)^T \mathcal{H}(t))^{-1} \mathcal{H}(t)^T$$

onto $N(t)$ to be continuously differentiable in t.

Conversely, given continuous projectors $Q(t)$, $P(t) = I - Q(t)$, consider a basis $\{d_1(t_0), d_2(t_0), \ldots, d_k(t_0)\}$ of $\operatorname{im} Q(t_0) = N(t_0)$ at any $t_0 \in \mathcal{J}$ and the differential equation

$$x'(t) = Q'(t)x(t).$$

It provides k linear independent C^1-solutions $d_1(\cdot), \ldots, d_k(\cdot)$ for the linear independent starting values $d_1(t_0), d_2(t_0), \ldots, d_k(t_0)$. Furthermore, given $P(t) := I - Q(t)$,

$$\begin{aligned}(P(t)d_j(t))' &= P'(t)d_j(t) + P(t)d_j'(t)\\ &= P'(t)d_j(t) + P(t)Q'(t)d_j(t)\\ &= P'(t)d_j(t) - P'(t)Q(t)d_j(t)\\ &= P'(t)(P(t)d_j(t))\end{aligned}$$

is satisfied for all $j = 1, \ldots, k$. Since $(Pd_j)(t_0) = 0$, the identity $(Pd_j)(t) = 0$ holds for all $t \in \mathcal{J}$. That is, $\{d_1(t), d_2(t), \ldots, d_k(t)\}$ forms a continuously differentiable basis of $N(t)$. □

Reflexive generalized inverse

Definition 2.7. *For a matrix $M \in L(\mathbb{R}^m, \mathbb{R}^n)$, a matrix $\tilde{M} \in L(\mathbb{R}^n, \mathbb{R}^m)$ is called a generalized inverse of M if*

$$\tilde{M}M\tilde{M} = \tilde{M}.$$

If the condition

$$M\tilde{M}M = M$$

holds as well, then $\tilde{M}$ is called a reflexive generalized inverse of M.

Observe that for any reflexive generalized inverse $\tilde{M}$ of M the matrices

$$(M\tilde{M})^2 = M\tilde{M}M\tilde{M} = M\tilde{M}, \qquad (\tilde{M}M)^2 = \tilde{M}M\tilde{M}M = \tilde{M}M$$

are projectors. Reflexive generalized inverses are not uniquely determined. Uniqueness is obtained if we require $M\tilde{M}$ and $\tilde{M}M$ to be *special* projectors. For instance, we could require them to satisfy

$$(M\tilde{M})^T = M\tilde{M}, \qquad (\tilde{M}M)^T = \tilde{M}M.$$

In this case $\tilde{M}$ is called the Moore-Penrose inverse of M, often denoted by M^+.

2.2 Weierstraß-Kronecker canonical form

A linear DAE with constant coefficients is a system of the form

$$Ex'(t) + Fx(t) = q(t), \quad t \in \mathcal{J}, \tag{2.5}$$

where E, $F \in L(\mathbb{R}^m)$ and $q(t) \in C(\mathcal{J}, \mathbb{R}^m)$, $\mathcal{J} \subseteq \mathbb{R}$. Thereby, the leading matrix E is singular. If the matrix E is nonsingular, the system (2.5) coincides with an

explicit regular linear ODE with constant coefficients. Solvability of the DAE (2.5) is closely related to the regularity of the matrix pair $\{E, F\}$, as illustrated below.

Definition 2.8. *The ordered pair of matrices $\{E, F\}$ forms a* regular matrix pencil *if the polynomial $p(\lambda) := \det(\lambda E + F)$ does not vanish identically. Otherwise, the pencil is called singular.*

Here, we exclude an equation of the form (2.5) with a singular matrix pencil $\{E, F\}$. This relies on the fact that the homogeneous system

$$Ex'(t) + Fx(t) = 0, \tag{2.6}$$

together with the initial condition $x(0) = 0$, for instance,

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

has an infinite-dimensional solution space if the matrix pencil is singular (see [27]).

A matrix pair $\{E, F\}$ with nonsingular E is always regular, and its polynomial p is of degree m. A regular matrix pencil $\{E, F\}$ can be transformed into the *Weierstraß-Kronecker canonical form* [21]. Namely, there exist nonsingular matrices L, $K \in L(\mathbb{R}^m)$ such that

$$LEK = \begin{bmatrix} I & \\ & \mathcal{N} \end{bmatrix}, \qquad LFK = \begin{bmatrix} W & \\ & I \end{bmatrix}, \tag{2.7}$$

where $W \in L(\mathbb{R}^{m-l})$ for some integer $1 \leq l \leq m$ and $\mathcal{N} \in L(\mathbb{R}^l)$ is a nilpotent Jordan block matrix with nilpotency index $\mu \leq l$, i.e., $\mathcal{N}^\mu = 0$, $\mathcal{N}^{\mu-1} \neq 0$. The integer μ is uniquely determined by the pair $\{E, F\}$. Proof of this result can be found in [21, 27]. The Kronecker index [21, 48] is defined in terms of the matrix pencil $\{E, F\}$.

Definition 2.9. *The Kronecker index μ of a regular pair $\{E, F\}$ with singular E is defined to be the nilpotency order μ in the Kronecker normal form* (2.7)*. We write ind$\{E, F\} = \mu$.*

Let us assume that the matrix pencil $\{E, F\}$ is regular, then the structure of the DAE (2.5) and (2.6) can be revealed via the Weierstraß-Kronecker canonical form. Premultiplying (2.5) by L and using the transformed variables $x = K \begin{bmatrix} y \\ z \end{bmatrix}$ the equivalent decoupled system reads

$$y'(t) + Wy(t) = p(t), \tag{2.8a}$$
$$\mathcal{N}z'(t) + z(t) = r(t), \quad t \in \mathcal{J}, \tag{2.8b}$$

with $Lq := \begin{bmatrix} p \\ r \end{bmatrix}$. Equation (2.8a) is an explicit linear constant coefficients ODE for the component y. Only for this component initial conditions may be imposed.

That means, this equation has $m - l$ dynamical degrees of freedom. The solution of (2.8b) is given by

$$z(t) \;=\; \sum_{j=0}^{\mu-1} (-1)^j \mathcal{N}^j r^{(j)}(t), \tag{2.9}$$

where μ denotes the index of the nilpotency of the Jordan block matrix $\mathcal{N}$ and r is assumed to be sufficiently smooth. It is clear that, for $\mu \geq 2$, equation (2.8b) may introduce differentiation problems; in order to calculate $z(t)$, some components of the right-hand side have to be differentiated $\mu - 1$ times. Only for $\mu = 1$ we have $\mathcal{N} = 0$, hence $z(t) = r(t)$ and no derivatives are involved in this case.

Definition 2.10. *A linear DAE* (2.5) *with constant coefficients is said to be regular or regular with Kronecker index* $\mu = ind\{E, F\}$ *if the pair* $\{E, F\}$ *is regular.*

In the case where E is nonsingular, the block $\mathcal{N}$ does not appear at all in the Weierstraß-Kronecker canonical form. Hence, this special case is categorized as a differential-algebraic problem with index $\mu = 0$. Obviously, an initial value problem for (2.5) only become solvable for *consistent initial conditions*

$$x(t_0) \;=\; K \begin{bmatrix} y(t_0) \\ z(t_0) \end{bmatrix} \;=\; K \begin{bmatrix} y_0 \\ z(t_0) \end{bmatrix},$$

where $y_0 \in \mathbb{R}^{m-l}$ is a free parameter, but $z(t_0)$ is completely determined from $r(t)$ via the relation (2.9).

In the homogeneous case (2.6), the explicit ODE (2.8a) reads

$$y'(t) + Wy(t) \;=\; 0,$$

while we obtain $z = 0$ in (2.8b), due to $r = 0$ and (2.9). Hence, the solution has the form

$$x(t) \;=\; K \begin{bmatrix} e^{-tW} \\ 0 \end{bmatrix} y_0, \quad y_0 \in \mathbb{R}^{m-l},$$

that means, the dimension of the solution space is $m - l$. Additionally, introducing the space

$$S_{EF} \;:=\; \{z \in \mathbb{R}^m : Fz \in \operatorname{im} E\},$$

implies that each solution of the homogeneous DAE (2.6) has to lie on this subspace, i.e., to satisfy $x \in S_{EF}$.

Unfortunately, the Kronecker index of a regular matrix pencil cannot be generalized to linear time-varying DAEs (or standard form linear DAEs) of the form

$$E(t)x'(t) + F(t)x(t) \;=\; q(t), \qquad t \in \mathcal{J}, \tag{2.10}$$

where $E(t)$, $F(t) \in C(\mathcal{J}, L(\mathbb{R}^m))$, $q(t) \in C(\mathcal{J}, \mathbb{R}^m)$ and $\mathcal{J} \subseteq \mathbb{R}$. The regularity of the matrix pair $\{E(t), F(t)\}$, for all $t \in \mathcal{J}$, does not guarantee the unique solvability of the relevant initial value problems.

Example 2.11. The DAE (2.10) given by the coefficients

$$E(t) = \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \quad F(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad q(t) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathcal{J} = \mathbb{R}, \tag{2.11}$$

reads

$$\begin{aligned} -tx_1'(t) + t^2 x_2'(t) + x_1(t) &= 0, \\ -x_1'(t) + tx_2'(t) + x_2(t) &= 0. \end{aligned}$$

Due to

$$\det(\lambda E(t) + F(t)) = (1 - \lambda t)(1 + \lambda t) + \lambda^2 t^2 = 1,$$

the matrix pencil $\{E(t), F(t)\}$ is regular for all $t \in \mathcal{J}$. It can be easily verified that x given by

$$x(t) = \gamma(t) \begin{bmatrix} t \\ 1 \end{bmatrix}$$

is a solution of the corresponding homogeneous initial value problem (2.11) together with $x(t_0) = 0$ for every $\gamma \in C^1(\mathcal{J}, \mathbb{R})$ with $\gamma(t_0) = 0$. In particular, there exists more than one solution.

Example 2.12. Consider the DAE

$$\begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix} x'(t) + \begin{bmatrix} 1 & -t \\ 0 & 0 \end{bmatrix} x(t) = q(t), \quad t \in \mathcal{J} = \mathbb{R}, \tag{2.12}$$

with $q \in C^2(\mathcal{J}, \mathbb{R}^2)$. Since

$$\det(\lambda E(t) + F(t)) = -\lambda t + \lambda t \equiv 0,$$

the matrix pair $\{E(t), F(t)\}$ is singular for all $t \in \mathcal{J}$. Now we write (2.12) as

$$\begin{aligned} x_1(t) - tx_2(t) &= q_1(t), \\ x_1'(t) - tx_2'(t) &= q_2(t). \end{aligned}$$

The first equation gives $x_1(t) = tx_2(t) + q_1(t)$. Differentiating this equation and inserting it into the second equation yields the unique solution

$$x_1(t) = q_1(t) + tq_2(t) - tq_1'(t), \quad x_2(t) = q_2(t) - q_1'(t).$$

Therefore, in this case, every initial value problem with consistent initial value is uniquely solvable. The corresponding homogeneous equation has only the trivial solution.

These observations induce many different approaches which generalize the Kronecker index of a regular matrix pencil to linear time-varying DAEs. One of these approaches is the *tractability index*, which is a projector-based method where the index notion is characterized via a sequence of matrix functions and subspaces. The solutions of DAEs can be explicitly described in terms of the original variables (see [27, 28, 57, 60, 61]). The computation of the Weierstraß-Kronecker canonical form is not required.

In order to apply the projector-based methods to linear time-varying DAEs of the form (2.10), see [58, 59, 60], we assume that the nullspace $N(t) := \ker E(t)$ is continuously differentiable for $t \in \mathcal{J}$. This means that there exists a C^1 projector function $Q(t)$ onto $N(t)$ and $P(t) := I - Q(t)$, $t \in \mathcal{J}$ (cf. Theorem 2.6 in Section 2.1). As shown in [60], since $E(t)Q(t) = 0$, we may insert $E(t) = E(t)P(t)$ into (2.10) and rewrite (2.10),

$$E(t)\{(Px)'(t) - P'(t)x(t)\} + F(t)x(t) = q(t),$$

or

$$E(t)(Px)'(t) + \{F(t) - E(t)P'(t)\}x(t) = q(t), \qquad t \in \mathcal{J}. \tag{2.13}$$

This makes it possible to seek for solutions within the function space

$$C_P^1(\mathcal{J}, \mathbb{R}^m) := \{x \in C(\mathcal{J}, \mathbb{R}^m) : Px \in C^1(\mathcal{J}, \mathbb{R}^m)\},$$

i.e., we are looking for continuous solutions that have continuously differentiable parts $(Px)(\cdot)$ and satisfy (2.10) pointwise.

Now, decomposing $x(t) = P(t)x(t) + Q(t)x(t)$ and denoting $B(t) := F(t) - E(t)P'(t)$ we transform (2.13) into

$$E(t)(Px)'(t) + B(t)P(t)x(t) + B(t)Q(t)x(t) = q(t),$$

and then into

$$\{E(t) + B(t)Q(t)\}\{P(t)(Px)'(t) + Q(t)x(t)\} + B(t)P(t)x(t) = q(t). \tag{2.14}$$

In the time-dependent setting, the (tractability) index-1 condition will be stated as the nonsingularity on $\mathcal{J}$ of the matrix

$$E_1(t) := E(t) + B(t)Q(t).$$

as stated in Lemma 2.4. In the following, we omit the argument t for notational simplicity. Hence, scaling (2.14) by E_1^{-1} implies

$$P(Px)' + E_1^{-1}BPx + Qx = E_1^{-1}q. \tag{2.15}$$

Premultiplying (2.15) by P and Q, respectively, and carrying out simple computations we obtain the decoupled system

$$\begin{aligned} (Px)' - P'Px + PE_1^{-1}BPx &= PE_1^{-1}q, \\ Qx + QE_1^{-1}BPx &= QE_1^{-1}q. \end{aligned}$$

Denoting u and v by $u := Px$ and $v := Qx$, we can rewrite this system as

$$u' - P'u + PE_1^{-1}Bu = PE_1^{-1}q, \tag{2.16a}$$
$$v = QE_1^{-1}q - QE_1^{-1}Bu. \tag{2.16b}$$

Similar to the approach based on the Kronecker index of a regular matrix pair, the system (2.16) provides a decoupling of the DAE (2.10) in terms of an explicit regular ODE (2.16a) for the component Px and an algebraic (i.e. derivative-free) equation (2.16b) for determining the component Qx. Equation (2.16a) is called the *inherent explicit regular ODE* of the linear time-varying index-1 DAE (2.10). The subspace $\operatorname{im} P(t)$, $t \in \mathcal{J}$, is an *invariant* subspace of the inherent explicit regular ODE (2.16a) in the sense that if a solution starts in $u(t_0) \in \operatorname{im} P(t_0)$, for some $t_0 \in \mathcal{J}$, it remains in the space $\operatorname{im} P(t)$, that is $u(t) \in \operatorname{im} P(t)$, for all $t \in \mathcal{J}$.

Consequently, for given $q \in C(\mathcal{J}, \mathbb{R}^m)$, $u_0 \in \operatorname{im} P(t_0)$, $x \in C_P^1(\mathcal{J}, \mathbb{R}^m)$ is a solution of (2.10) if and only if it can be written as

$$x(t) = u(t) + v(t),$$

where $u(t) \in C^1(\mathcal{J}, \mathbb{R}^m)$ is a solution of (2.16a) in the invariant space $\operatorname{im} P$ and $v(t) \in C(\mathcal{J}, \mathbb{R}^m)$ is explicitly given by (2.16b).

The linear time-varying DAEs (2.10) can be reformulated to a DAE of the form (2.17), see below, by denoting $D(t) = P(t)$, $B(t) := F(t) - E(t)P'(t)$. Hence, results obtained for the system (2.17) can be applied in particular to the standard form (2.10), see [2, 41, 63, 64, 65, 66]. This is why we don't give more details of this decoupling procedure for the higher index cases of linear time-varying DAEs (2.10). For additional details concerning DAEs (2.10) with higher index we refer to [60].

2.3 Linear DAEs with properly stated leading term

Consider linear time-varying DAEs of the form

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \qquad t \in \mathcal{J}, \tag{2.17}$$

where $A \in C(\mathcal{J}, L(\mathbb{R}^n, \mathbb{R}^m))$, $D \in C(\mathcal{J}, L(\mathbb{R}^m, \mathbb{R}^n))$, $B \in C(\mathcal{J}, L(\mathbb{R}^m))$, $q \in C(\mathcal{J}, \mathbb{R}^m)$, and $\mathcal{J} \subseteq \mathbb{R}$ is an interval.

The leading term $A(t)(D(t)x(t))'$ specifies precisely which components of the solution need to be differentiated. The form of the leading term in (2.17) has been motivated by the study of linear DAEs and their adjoint equations. In addition, this formulation arises in various control and circuit applications. See [1, 2, 3, 41, 50, 54, 62, 63, 64]. As discussed in [77] the properly stated leading term in (2.17) provides precise input-output functional descriptions of linear time-varying DAEs.

Definition 2.13. *A continuous function $x(\cdot) : \mathcal{J} \to \mathbb{R}^m$ is said to be a solution of* (2.17) *on $\mathcal{J} \subseteq \mathbb{R}$ if $Dx \in C^1(\mathcal{J}, \mathbb{R}^n)$ and* (2.17) *is satisfied for all $t \in \mathcal{J}$. Let*

$$C_D^1(\mathcal{J}, \mathbb{R}^m) := \{x \in C(\mathcal{J}, \mathbb{R}^m) : Dx \in C^1(\mathcal{J}, \mathbb{R}^n)\}$$

denote the corresponding function space.

In contrast to P_E in (1.4) or P in (2.13), neither A nor D in (2.17) needs to be a projector. Nevertheless, they have to be well matched in the sense of the following Definition 2.14.

Definition 2.14. *The leading term of the DAE* (2.17) *is said to be* properly stated on the interval $\mathcal{I} \subseteq \mathcal{J}$ *if the coefficients $A(t)$ and $D(t)$ are well matched so that the decomposition*

$$\ker A(t) \oplus \operatorname{im} D(t) = \mathbb{R}^n, \quad t \in \mathcal{I}, \tag{2.18}$$

holds and there exists a projector function $R \in C^1(\mathcal{I}, L(\mathbb{R}^n))$ such that

$$\ker R(t) = \ker A(t), \quad \operatorname{im} R(t) = \operatorname{im} D(t), \quad t \in \mathcal{I}.$$

By definition, if the leading term is properly stated on $\mathcal{I} \subseteq \mathcal{J}$, then the matrix functions A, D and the product AD have a common constant rank on $\mathcal{I}$ [63]. The continuously differentiable projector function R onto $\operatorname{im} D(t)$ along $\ker A(t)$ exists, for all $t \in \mathcal{I}$, when both $\ker A(t)$ and $\operatorname{im} D(t)$ are C^1 spaces, i.e., they have constant dimension, are spanned by C^1 basis functions, and the transversality condition (2.18) holds.

2.3.1 Matrix chain and admissible projectors

Following [64], we construct a sequence of matrix functions and subspaces for (2.17) to define the tractability index. For the sake of simplicity, we drop the argument t in the following considerations, where all relations are meant pointwise for $t \in \mathcal{I}$. Let the DAE (2.17) have a properly stated leading term on the interval $\mathcal{I} \subseteq \mathcal{J} \subseteq \mathbb{R}$.

Since the matrix function D is continuous with constant rank, we may choose projector functions Q_0, P_0, $\Pi_0 \in C(\mathcal{I}, L(\mathbb{R}^m))$ such that

$$Q_0^2 = Q_0, \quad \operatorname{im} Q_0 = \ker D, \quad \Pi_0 = P_0 = I - Q_0. \tag{2.19}$$

In addition, we take the projector functions $P_0(t)$ and $R(t)$ to determine the reflexive generalized inverse $D^-(t) \in C(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^m))$ uniquely. Hence, $D^-(t)$ is the reflexive generalized inverse of $D(t)$ defined by

$$DD^-D = D, \quad D^-DD^- = D^-, \quad DD^- = R, \quad D^-D = P_0, \tag{2.20}$$

where the dependence in t is omitted. If there is another reflexive generalized

inverse $\tilde{D}^-$ satisfying (2.20), then

$$\tilde{D}^- = \tilde{D}^- D\tilde{D}^- = \tilde{D}^- R = \tilde{D}^- DD^- = P_0 D^- = D^- DD^- = D^-.$$

Introduce further

$$G_0 := AD, \quad N_0 := \ker D = \ker G_0, \quad B_0 := B, \tag{2.21}$$

and, for $i \geq 0$, as long as the expressions exist,

$$G_{i+1} = G_i + B_i Q_i, \quad N_{i+1} := \ker G_{i+1}, \tag{2.22}$$

choose projector functions P_{i+1}, Q_{i+1} such that

$$Q_{i+1}^2 = Q_{i+1}, \quad \operatorname{im} Q_{i+1} = N_{i+1}, \quad P_{i+1} = I - Q_{i+1}, \tag{2.23}$$

and define

$$\begin{aligned} \Pi_{i+1} &:= \Pi_i P_{i+1}, \\ B_{i+1} &= B_i P_i - G_{i+1} D^- (D\Pi_{i+1} D^-)' D\Pi_i. \end{aligned} \tag{2.24}$$

The abbreviations Π_i, $i \geq 1$, stand for the product $P_0 \cdots P_i$.

Up to now, the matrix function $Q_i(t) \in L(\mathbb{R}^m)$, $i \geq 0$, is defined to be *any* projector onto the nullspace of $G_i(t)$, $t \in \mathcal{I}$, and P_i is its complementary projector. This implies that the matrix function G_i depends on how the projectors Q_0 (resp. D^-) and Q_i, $i \geq 1$, are chosen. However, for the purpose of decoupling the DAE (2.17) later, it is useful to restrict the variety of possible projector functions Q_i and to choose a so-called *admissible* projector function [65, 66].

Definition 2.15. *Let* (2.17) *be a DAE with proper leading term on $\mathcal{I} \subseteq \mathcal{J}$. The projector functions $Q_0, \ldots, Q_k$ with $k \in \mathbb{N}$, are said to be* admissible *(for the DAE* (2.17) *on $\mathcal{I}$), if the following properties hold for $i = 0, 1, \ldots, k$:*

(a) G_i has constant rank r_i on $\mathcal{I}$,

(b) $N_i = \ker G_i$, $i \geq 1$, satisfies the conditions

$$\widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i = \{0\}, \tag{2.25}$$

$$N_0 + \cdots + N_{i-1} \subseteq \ker Q_i, \tag{2.26}$$

(c) Π_i is continuous and $D\Pi_i D^-$ is continuously differentiable.

If the projector functions $Q_0, \ldots, Q_k$ are admissible, then the corresponding matrix function sequence (2.19)-(2.24) *is said to be admissible up to level k.*

In the light of Definition 2.15, any continuous projector Q_0 onto $N_0 = \ker G_0$ is admissible for a properly stated DAE (2.17), since G_0 has on $\mathcal{I}$ constant rank $r_0 = r$ and $DP_0D^- = DD^- = R$ is continuously differentiable.

As shown in [64], the trivial intersections $\widehat{N}_i$, $i \geq 1$, as specified in (2.25), make it possible to choose the continuous projector Q_i onto $N_i = \ker G_i$ in a way such that

$$N_0 + \cdots + N_{i-1} \subseteq \ker Q_i.$$

Since Q_j projects onto $N_j = \ker G_j$ for $0 \leq j < i$, it follows that

$$Q_i Q_j = 0. \tag{2.27}$$

Observe that a certain nontrivial intersection $N_{i*+1} \cap N_{i*}$ would yield the whole matrix sequence $\{G_k\}_{k\geq 0}$ to consist of singular matrices only [64].

The property (b) ensures that certain products of projector functions Π_1, ..., Π_i, P_0Q_1, ..., $\Pi_{i-1}Q_i$ are again projectors. All terms in the decomposition of the C^1 projector function R belong also to C^1 (see e.g. [64]). These are

$$R = DD^- = DP_0D^- = D\Pi_i D^- + D\Pi_{i-1}Q_iD^- + \cdots + DP_0Q_1D^-.$$

It guarantees also the existence of the derivatives in the expression for B_{i+1} in (2.24).

Proposition 2.16. *Let Q_0,..., Q_k be admissible projector functions. Then*

$$\ker \Pi_i = N_0 + \cdots + N_i, \quad i = 0, \ldots, k.$$

Proof : We verify this assertion by induction. Let $\ker P_0 = N_0$. $P_0P_1z = 0$ implies $z_0 := (I - Q_1)z \in N_0$, hence $z = Q_1z + z_0 \in N_0 + N_1$. On the other hand, due to $z \in N_0 + N_1$, we may decompose each z as $z = Q_0w_0 + Q_1w_1$. Then we compute $P_0P_1z = P_0P_1Q_0w_0 = P_0(I - Q_1)Q_0w_0 = 0$. Therefore $N_0 + N_1 \subseteq \ker P_0P_1$.
For induction, let $\ker P_0 \cdots P_{i-1} = N_0 + \cdots + N_{i-1}$ for $i \leq k$. From $P_0 \cdots P_iz = 0$, i.e. $P_iz \in \ker P_0 \cdots P_{i-1}$, we define $\tilde{z} := (I - Q_i)z \in N_0 + \cdots + N_{i-1}$ and find $z = \tilde{z}+Q_iz \in N_0+\cdots+N_i$. Conversely, for $z \in N_0+\cdots+N_i = (N_0+\cdots+N_{i-1})+N_i$ we decompose z as $z = z_* + z_i$ with $z_* \in N_0 + \cdots + N_{i-1}$ and $z_i \in N_i$. Since $N_0 + \cdots + N_{i-1} \subseteq \ker Q_i$, we have $Q_iz_* = 0$ and hence $z_* = P_iz_*$. Now we compute $P_0 \cdots P_iz = P_0 \cdots P_{i-1}P_iz_* + P_0 \cdots P_{i-1}P_iz = P_0 \cdots P_{i-1}z_* = 0$. Thus, $z \in \ker P_0 \cdots P_{i-1}$. □

The idea of tractability index is to replace the smoothness requirements for the coefficients by the requirement that they have to be smooth on certain subspaces, e.g. on $\ker E(t)$ in standard form (2.10). If there exists a minimal nonnegative index i for which the matrix G_i is nonsingular on the entire interval $\mathcal{I}$, the linear DAE (2.17) is said to be regular as formally indicated below.

Definition 2.17. *The DAE* (2.17) *with a properly stated leading term on $\mathcal{I} \subseteq \mathcal{J}$ is said to be*

(a) regular with tractability index zero *(on $\mathcal{I}$) if both A and D are nonsingular on $\mathcal{I}$,*

(b) regular with tractability index $\mu \geq 1$ *(on $\mathcal{I}$) if there exist admissible projectors* $Q_0, \ldots, Q_{\mu-1}$ *such that* $0 \leq r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m$,

(c) regular *(on $\mathcal{I}$) if there is an integer μ such that the DAE* (2.17) *is regular with tractability index μ.*

The numbers $r_0, \ldots, r_\mu$ as well as μ and $d = m - \sum_{i=0}^{\mu-1}(m - r_i)$, defined by the matrix function sequence (2.19)-(2.24)*, are called* characteristic values *of a regular DAE* (2.17)*.*

Definition 2.18. *A point $t_* \in \mathcal{J}$ is said to be regular for* (2.17)*, if there exists an open interval $\mathcal{I} \subseteq \mathcal{J}$ of t_* such that the DAE* (2.17) *is regular on $\mathcal{I}$ according to Definition 2.17. The interval $\mathcal{I}$ will be called the* regularity interval *for t_* and the union of the regularity intervals within $\mathcal{J}$ will be denoted by $\mathcal{J}_{reg}$. A point $t_* \in \mathcal{J} - \mathcal{J}_{reg}$ is said to be* critical*.*

As acknowledged in Proposition 2.10 of [64], neither the definition of a regular DAE nor the characteristic values $r_0, r_1, \ldots, r_{\mu-1}, r_\mu = m$, μ, d are dependent on the special choice of the admissible projector functions Q_0, …, $Q_{\mu-1}$.

Note that the transversality and smoothness conditions in Definition 2.15 become trivial at level μ. Remark also that a constant-coefficient DAE is regular with tractability index μ if and only if the matrix pencil $\{G_0, B\}$ is regular with Kronecker index μ [64].

We will complete this subsection with some well-known classes of DAEs and show how these systems correspond to the DAEs in the properly stated framework. The first system is the so-called *semi-explicit* DAEs playing the important role in system modeling. These systems arise in applications when algebraic constraints are explicitly added to a set of differential relations. In turn, the well understood *Hessenberg* DAEs having index 2 will be stated in Example 2.20. Hessenberg form systems occur in mechanics and, particalarly, in electrical circuit theory. The last example of this subsection is related to Example 1.3 from the introduction.

Example 2.19. We consider the index-1 semi-explicit DAE of the form

$$x_1'(t) + B_{11}(t)x_1(t) + B_{12}(t)x_2(t) = q_1(t), \tag{2.28a}$$
$$B_{21}(t)x_1(t) + B_{22}(t)x_2(t) = q_2(t), \tag{2.28b}$$

where $B_{22}(t)$ is supposed to be nonsingular on the interval $\mathcal{J}$. Letting $m = m_1 + m_2$, $n = m_1$ and taking

$$A = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} I & 0 \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \quad D^- = \begin{bmatrix} I \\ 0 \end{bmatrix},$$

we may write (2.28) in the form (2.17) with properly stated leading term. Then we have

$$G_0 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \quad G_1 = \begin{bmatrix} I & B_{12} \\ 0 & B_{22} \end{bmatrix},$$

with $r_0 = m_1$, $r_1 = m$. The nonsingularity of B_{22} makes G_1 nonsingular and the problem is index 1 in our context (cf. Definition 2.17). Let us notice that this system does not contain any critical point, cf. Definition 3.5 in Chapter 3, since the matrix functions A and G_0 have constant rank.

Example 2.20. Consider an index-2 DAE in Hessenberg form

$$x_1'(t) + B_{11}(t)x_1(t) + B_{12}(t)x_2(t) = q_1(t), \tag{2.29a}$$
$$B_{21}(t)x_1(t) = q_2(t), \tag{2.29b}$$

where the product $B_{21}B_{12} \in L(\mathbb{R}^{m_2})$ is nonsingular on $\mathcal{J}$. Observe that this system is obtained from the problem (2.28) if $B_{22} = 0$. This system can be written in form (2.17) with properly stated leading term by setting

$$A = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} I & 0 \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

The matrix functions G_0, Q_0, and G_1 are

$$G_0 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \quad G_1 = \begin{bmatrix} I & B_{12} \\ 0 & 0 \end{bmatrix}.$$

The matrix function G_1 has rank $r_1 = m_1$ and a nontrivial nullspace. Let $\mathcal{H}(t)$ be a projector function onto $\operatorname{im} B_{12}$, and B_{12}^- be a reflexive generized inverse such that $B_{12}B_{12}^- = \mathcal{H}(t)$, $B_{12}^-B_{12} = I$. Then the projector function Q_1

$$Q_1 = \begin{bmatrix} \mathcal{H} & 0 \\ -B_{12}^- & 0 \end{bmatrix}$$

verifies $Q_1Q_0 = 0$. This leads to projector functions

$$P_0P_1 = \Pi_1 = \begin{bmatrix} I - \mathcal{H} & 0 \\ 0 & 0 \end{bmatrix}, \quad P_0Q_1 = \begin{bmatrix} \mathcal{H} & 0 \\ 0 & 0 \end{bmatrix}, \quad D\Pi_1 D^- = I - \mathcal{H}.$$

We compute

$$B_1 = \begin{bmatrix} B_{11} & 0 \\ B_{21} & 0 \end{bmatrix} - \begin{bmatrix} -\mathcal{H}' & 0 \\ 0 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} I + (B_{11} + \mathcal{H}')\,\mathcal{H} & B_{12} \\ B_{21}\mathcal{H} & 0 \end{bmatrix},$$

and check the nullspace of G_2. Let $z \in \ker G_2$, then

$$z_1 + (B_{11} + \mathcal{H}')\,\mathcal{H}z_1 + B_{12}z_2 = 0, \quad B_{21}\mathcal{H}z_1 = 0.$$

The regularity of $B_{21}B_{12}$ yields $B_{21}\mathcal{H}z_1 = B_{21}B_{12}B_{12}^-z_1 = 0$, and hence $B_{12}^-z_1 = 0$ meaning that $\mathcal{H}z_1 = B_{12}B_{12}^-z_1 = 0$ in the second equation. Now, the first equation simplifies to $z_1 + B_{12}z_2 = 0$. Multiplying by B_{12}^- we obtain $z_2 = 0$, and then $z_1 = 0$. Therefore, the matrix function G_2 is nonsingular and this problem is regular with tractability index 2. Note that, as in Example 2.19, the critical points according to Definition 3.5 cannot arise in the index-2 Hessenberg DAEs, because the matrix functions A, G_0, and G_1 have constant rank.

Example 2.21. As introduced in Example 1.3, Equation (1.10) can be written as a DAE with proper formulated leading term by choosing

$$A = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & \eta t \end{bmatrix}, \quad B = \begin{bmatrix} 1 & \eta t \\ 0 & 1 \end{bmatrix}, \quad D^- = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The matrix $G_0 = AD$ and its nullspace read

$$G_0 = \begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix}, \quad N_0 = \{z \in \mathbb{R}^2 : z_1 + \eta t z_2 = 0\}.$$

Taking

$$Q_0 = \begin{bmatrix} 0 & -\eta t \\ 0 & 1 \end{bmatrix}, \quad \Pi_0 = \begin{bmatrix} 1 & \eta t \\ 0 & 0 \end{bmatrix},$$

we obtain

$$G_1 = \begin{bmatrix} 0 & 0 \\ 1 & 1 + \eta t \end{bmatrix}, \quad N_1 = \{z \in \mathbb{R}^2 : z_1 + (1 + \eta t) z_2 = 0\}.$$

An admissible choice for the projector Q_1 verifying $Q_1 Q_0 = 0$, and $P_1 = I - Q_1$ is

$$Q_1 = \begin{bmatrix} 1 + \eta t & \eta t(1 + \eta t) \\ -1 & -\eta t \end{bmatrix}, \quad P_1 = \begin{bmatrix} -\eta t & -\eta t(1 + \eta t) \\ 1 & 1 + \eta t \end{bmatrix},$$

which yields

$$P_0 P_1 = \Pi_1 = 0, \quad P_0 Q_1 = \begin{bmatrix} 1 & \eta t \\ 0 & 0 \end{bmatrix}, \quad D\Pi_1 D^- = 0.$$

Then we compute

$$B_1 = B_0 P_0 = \begin{bmatrix} 1 & \eta t \\ 0 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 1 & \eta t \\ 1 & 1 + \eta t \end{bmatrix}.$$

Due to $\det G_2 = 1$ the DAE (1.10) is regular with tractability index 2.

Remark 2.22. *Equivalently, the DAE* (2.17) *with properly stated leading term is regular on $\mathcal{I}$ with tractability index $\mu \geq 1$, if there are admissible projector functions such that the matrix functions $G_i, 0 \leq i < \mu$, defined in* (2.19)-(2.24)*, are singular and G_μ is nonsingular for all $t \in \mathcal{I}$.*

The definition of B_{i+1} presented here and taken from [59, 60] *is different from $\mathcal{B}_{i+1} := \mathcal{B}_i P_i$ in* [2, 43] *concerning the DAEs of at most index 2. There, the matrix function $\mathcal{G}_2$ is given by*

$$\mathcal{G}_2 := G_1 + B_0 P_0 Q_1.$$

Due to $G_2 P_1 = G_1$, G_2 may be written as a product

$$\begin{aligned} G_2 &= (G_1 + B_0 P_0 Q_1)(I - P_1 D^- (D\Pi_1 D^-)' D\Pi_0 Q_1) \\ &= \mathcal{G}_2 (I - P_1 D^- (D\Pi_1 D^-)' D\Pi_0 Q_1). \end{aligned}$$

Since the second factor of this product is nonsingular, it holds that

$$\operatorname{rank} G_2 = \operatorname{rank} \mathcal{G}_2.$$

2.3.2 Decoupling for regular DAEs

By means of the projector-based method, a linear DAE can be decomposed into three parts: the dynamic and the algebraic parts and the parts describing the inherent differentiation problem (cf. [64, 65, 76]). This allows to investigate the structure of a DAE. In particular, the decoupling procedure plays a key role in the study of stability and convergence properties of numerical methods which is carried out in Chapter 6.

Let the leading term of the DAE (2.17) be properly stated on $\mathcal{I} \subseteq \mathcal{J}$. As indicated above, we can define a continuous projector Q_0 onto the space $\ker AD$, the corresponding projector $\Pi_0 = P_0 = I - Q_0$, as well as the generalized reflexive inverse D^-. Since $R = DD^-$ is a projector along $\ker A$, the leading matrix function A in (2.17) can be written as $A = AR = ADD^- = G_0D^-$. Thus, (2.17) can be transformed into

$$G_0D^-(Dx)' + Bx = q. \tag{2.30}$$

Due to $x = P_0x + Q_0x = \Pi_0x + Q_0x$, this gives

$$G_0D^-(Dx)' + B\Pi_0x + BQ_0x = q.$$

Since $G_0D^- = G_1\Pi_0D^- = G_1D^-$ and $BQ_0 = (G_0 + BQ_0)Q_0 = G_1Q_0$, we have

$$G_1D^-(Dx)' + B\Pi_0x + G_1Q_0x = q, \tag{2.31}$$

For the moment, we assume that DAE (2.17) is regular on $\mathcal{I}$ with tractability index 1, i.e., that G_1 is nonsingular on $\mathcal{I}$. Then we can scale (2.31) by G_1^{-1} and obtain

$$D^-(Dx)' + G_1^{-1}BD^-Dx + Q_0x = G_1^{-1}q. \tag{2.32}$$

Multiplication of (2.32) by $D\Pi_0$ and Q_0, respectively, yields the system

$$DD^-u' + D\Pi_0G_1^{-1}BD^-u = D\Pi_0G_1^{-1}q, \tag{2.33a}$$

$$v_0 = Q_0G_1^{-1}q - Q_0G_1^{-1}B\Pi_0D^-u, \tag{2.33b}$$

where $u := Dx$, $v_0 := Q_0x$.

Using the product rule $DD^-u' = (DD^-Dx)' - (DD^-)'u = (Dx)' - (D\Pi_0D^-)'u$, (2.33a) can be transformed into

$$u' - (D\Pi_0D^-)'u + D\Pi_0G_1^{-1}BD^-u = D\Pi_0G_1^{-1}q. \tag{2.34}$$

Equation (2.34) represents an explicit ODE determining the component $u = Dx$ and is called the *inherent explicit regular ODE* of the index-1 DAE (2.17). Equation

(2.33b) is an algebraic equation for determining the component $v_0 = Q_0 x$.

Consequently, $x \in C_D^1(\mathcal{I}, \mathbb{R}^m)$ is a solution of (2.17) if it can be expressed as

$$x = \Pi_0 x + Q_0 x = D^- u + v_0 = (I - Q_0 G_1^{-1} B \Pi_0) D^- u + Q_0 G_1^{-1} q,$$

where u is a C^1 solution of (2.34) in the invariant space $\operatorname{im} D$, and v_0 is given by (2.33b). For the invariant property of the space $\operatorname{im} D$ one can verify as in the case of linear time-dependent DAEs (2.10). Recall, e.g. from [27], that $(I - Q_0 G_1^{-1} B P_0)$ is nonsingular and $(I - Q_0 G_1^{-1} B P_0) P_0$ is the canonical projector onto $S_0 = \{z \in \mathbb{R}^m : Bz \in \operatorname{im} G_0\}$ along the subspace N_0.
On the other hand, if $u \in C_D^1(\mathcal{I}, \mathbb{R}^n)$, $v_0 \in C(\mathcal{I}, \mathbb{R}^m)$ satisfy (2.33) with $u(t_0) \in \operatorname{im} D(t_0)$, then

$$x \;=\; D^- u + v_0$$

is a solution of the original DAE (2.17).

This decoupling procedure can be extended to DAEs with arbitrary index. Let (2.17) be a regular DAE with tractability index μ on $\mathcal{I}$. Due to $G_0 = G_1 P_0 = G_2 P_1 P_0 = \cdots = G_\mu P_{\mu-1} \cdots P_1 P_0$, we can transform (2.17) into

$$G_\mu P_{\mu-1} \cdots P_1 P_0 D^- (Dx)' + Bx = q. \tag{2.35}$$

Writing $B = B_0 = B_0 \Pi_{\mu-1} + B_0 \Pi_{\mu-2} Q_{\mu-1} + \ldots + B_0 \Pi_0 Q_1 + B_0 Q_0$, taking into account the definition of B_i, and using the relations $B_i Q_i = G_\mu Q_i$, $G_i = G_\mu P_{\mu-1} \cdots P_i$, and

$$\begin{aligned} B_0 \Pi_{i-1} Q_i &= B_i Q_i + \sum_{j=1}^{i} G_j D^- (D \Pi_j D^-)' D \Pi_{i-1} Q_i \\ &= G_\mu Q_i + \sum_{j=1}^{i} G_\mu P_{\mu-1} \cdots P_j D^- (D \Pi_j D^-)' D \Pi_{i-1} Q_i, \end{aligned}$$

for $i = 1, \ldots, \mu - 1$, equation (2.35) is equivalent to

$$\begin{aligned} G_\mu P_{\mu-1} \cdots P_1 P_0 D^- (Dx)' + B_0 \Pi_{\mu-1} x + \sum_{j=0}^{\mu-1} G_\mu Q_j x \\ + \sum_{i=1}^{\mu-1} \sum_{j=1}^{i} G_\mu P_{\mu-1} \cdots P_j D^- (D \Pi_j D^-)' D \Pi_{i-1} Q_i x = q. \end{aligned} \tag{2.36}$$

Scaling by G_μ^{-1} we derive the equation

$$\begin{aligned} P_{\mu-1} \cdots P_1 P_0 D^- (Dx)' + G_\mu^{-1} B_0 \Pi_{\mu-1} x + \sum_{j=0}^{\mu-1} Q_j x \\ + \sum_{i=1}^{\mu-1} \sum_{j=1}^{i} P_{\mu-1} \cdots P_j D^- (D \Pi_j D^-)' D \Pi_{i-1} Q_i x = G_\mu^{-1} q. \end{aligned} \tag{2.37}$$

This equation coincides, for $\mu = 1$, with (2.32) since $P_0 D^- = D^- D D^- = D^-$ and $\Pi_0 = P_0 = D^- D$. The derivatives $(D\Pi_j D^-)'$ start to appear for $\mu = 2$. The symbol Π_j is used again to stand for the product $P_0 \cdots P_j$.

As shown in [64, 65], according to the decomposition

$$I = \Pi_{\mu-1} + Q_0 P_1 \cdots P_{\mu-1} + Q_1 P_2 \cdots P_{\mu-1} + \cdots + Q_{\mu-2} P_{\mu-1} + Q_{\mu-1} \tag{2.38}$$

one can split (2.37) into $\mu + 1$ separate equations corresponding to the terms involved in (2.38). Namely, premultiplying (2.37) by $D\Pi_{\mu-1}$ and making use of $D\Pi_{\mu-1} D^- (Dx)' = (D\Pi_{\mu-1} x)' - (D\Pi_{\mu-1} D^-)' Dx$ yield the inherent explicit regular ODE for the component $u := D\Pi_{\mu-1} x$ as depicted in (2.40). Further, multiplying (2.37) by $Q_{\mu-1}$ gives

$$Q_{\mu-1} x + Q_{\mu-1} G_\mu^{-1} B_0 \Pi_{\mu-1} x = Q_{\mu-1} G_\mu^{-1} q.$$

If $\mu \geq 2$, we multiply once again by $\Pi_{\mu-2}$, so that the expression (2.41a) results. In turn, multiplying (2.37) by $Q_0 P_1 \cdots P_{\mu-1}$ and by $\Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1}$ for $i = 1, \ldots, \mu-2$, we obtain the explicit algebraic relations of the components $v_0 := Q_0 x$ and $v_i := \Pi_{i-1} Q_i x$, $i = 1, \ldots, \mu-1$, as shown in (2.41b) below.

Theorem 2.23. *Let* (2.17) *be a regular DAE with tractability index* μ *on* $\mathcal{I} \in \mathcal{J}$. *Then* $x \in C_D^1(\mathcal{I}, \mathbb{R}^m)$ *solves* (2.17) *if and only if it can be written as*

$$x \;=\; D^- u + v_0 + \cdots + v_{\mu-1}, \tag{2.39}$$

where $u = D\Pi_{\mu-1} x \in C^1(\mathcal{I}, \mathbb{R}^n)$ *is a solution of the inherent explicit regular ODE*

$$u' - (D\Pi_{\mu-1} D^-)' u + D\Pi_{\mu-1} G_\mu^{-1} B D^- u \;=\; D\Pi_{\mu-1} G_\mu^{-1} q, \tag{2.40}$$

lying on the invariant space $\operatorname{im} D\Pi_{\mu-1}$, *while the components* $v_0 = Q_0 x \in C(\mathcal{I}, \mathbb{R}^m)$ *and* $v_i = \Pi_{i-1} Q_i x \in C_D^1(\mathcal{I}, \mathbb{R}^m)$, $i = 1, \ldots, \mu-1$, *satisfy*

$$v_{\mu-1} = \mathcal{L}_{\mu-1} q - \mathcal{K}_{\mu-1} D^- u, \tag{2.41a}$$

$$v_i = \mathcal{L}_i q - \mathcal{K}_i D^- u + \sum_{j=i+1}^{\mu-1} \mathcal{N}_{ij} (D v_j)' + \sum_{j=i+2}^{\mu-1} \mathcal{M}_{ij} v_j, \; i = 0, \ldots, \mu-2. \tag{2.41b}$$

The continuous coefficients $\mathcal{L}_{\mu-1}$, $\mathcal{K}_{\mu-1}$ *in* (2.41a) *read*

$$\mathcal{L}_{\mu-1} \;=\; Q_{\mu-1} G_\mu^{-1}, \qquad \mathcal{K}_{\mu-1} = Q_{\mu-1} G_\mu^{-1} B \Pi_{\mu-1} \qquad \text{for } \mu = 1,$$

$$\mathcal{L}_{\mu-1} \;=\; \Pi_{\mu-2} Q_{\mu-1} G_\mu^{-1}, \quad \mathcal{K}_{\mu-1} = \Pi_{\mu-2} Q_{\mu-1} G_\mu^{-1} B \Pi_{\mu-1} \quad \text{for } \mu \geq 2.$$

The continuous coefficients $\mathcal{L}_i$, $\mathcal{K}_i$, $\mathcal{N}_{ij}$, *and* $\mathcal{M}_{ij}$ *in* (2.41b) *are given by*

$$\mathcal{L}_0 \;=\; Q_0 P_1 \cdots P_{\mu-1} G_\mu^{-1}, \quad \mathcal{L}_i \;=\; \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} G_\mu^{-1}, \quad i = 1, \ldots, \mu-2,$$

$$
\begin{aligned}
\mathcal{N}_{01} &= Q_0Q_1D^-, \quad \mathcal{N}_{0j} = Q_0P_1\cdots P_{j-1}Q_jD^-, \quad j=2,\dots,\mu-1,\\
\mathcal{N}_{i,i+1} &= \Pi_{i-1}Q_iQ_{i+1}D^-, \quad i=1,\dots,\mu-2,\\
\mathcal{N}_{ij} &= \Pi_{i-1}Q_iP_{i+1}\cdots P_{j-1}Q_jD^-, \quad j=i+2,\dots,\mu-1, \quad i=1,\dots,\mu-2,\\
\mathcal{K}_0 &= Q_0P_1\cdots P_{\mu-1}G_\mu^{-1}B\Pi_{\mu-1} + Q_0P_1\cdots P_{\mu-1}P_0D^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1},\\
\mathcal{K}_i &= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}B\Pi_{\mu-1}\\
&\quad + \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}P_iD^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}, \quad i=1,\dots,\mu-2,\\
\mathcal{M}_{ij} &= -\Pi_{i-1}Q_i\{Q_{i+1}D^-(D\Pi_iQ_{i+1}D^-)' + P_{i+1}Q_{i+2}D^-(D\Pi_{i+1}Q_{i+2}D^-)'\\
&\quad + \cdots + P_{i+1}\cdots P_{\mu-2}Q_{\mu-1}D^-(D\Pi_{\mu-2}Q_{\mu-1}D^-)'\}D\Pi_{j-1}Q_j\\
&\quad - \sum_{l=1}^{j}\Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}P_{\mu-1}\cdots P_lD^-(D\Pi_lD^-)'D\Pi_{j-1}Q_j,\\
&\qquad j=i+2,\dots,\mu-1, \quad i=1,\dots,\mu-2.
\end{aligned}
$$

Proof of this result can be found in [64, 65]. For $\mu = 2$ we refer to Chapter 5.

Remark 2.24. *In* [66] *a particular choice of the admissible projectors* $Q_0,\dots,Q_{\mu-1}$ *yields a so-called fine decoupling with vanishing coefficients* $\mathcal{K}_1,\dots,\mathcal{K}_{\mu-1}$. *Thereby,* Q_0 *can be chosen arbitrarily. Also, it is possible to construct admissible projector functions* $Q_0,\dots,Q_{\mu-1}$ *in such a way that the coefficient* $\mathcal{K}_0 = 0$ *resulting a complete decoupling.*

Chapter 3

Critical Points of DAEs

In this chapter, we consider DAEs where some of the assumptions under which the regularity of DAE is defined fail at certain - so called *critical* - points. Roughly speaking, we may associate with critical points a non-existence or a non-uniqueness of DAE solutions. In particular, a critical point may arise if the matrix function G_i (from the matrix chain (2.19)-(2.24)) is allowed to change its rank. Within the tractability index framework, a critical point makes it impossible to choose admissible projector functions satisfying Definition 2.15 to support the regularity of linear DAEs.

In [69, 70] critical points of the linear DAE (2.17), which will be addressed in this chapter, have been classified in cases when the algebraic constant-rank and transversality conditions do not hold. Recall that the regular set $\mathcal{J}_{reg}$ is defined as the union of the regularity intervals.

Definition 3.1. *Assume that the DAE* (2.17) *has continuous coefficients* $A(t)$, $D(t)$, $B(t)$. *A point* t_* *is said to be* critical *if there is no regularity interval containing it.*

Definition 3.2. *A continuous matrix function* $G : \mathcal{I} \to L(\mathbb{R}^k)$, *where* $\mathcal{I} \subseteq \mathcal{J} \subseteq \mathbb{R}$ *is an interval, has a* rank drop *at* $t_* \in \mathcal{I}$, *if each neighborhood of* t_* *contains points where the rank of* G *is different from* $\operatorname{rank} G(t_*)$. *Then,* t_* *is called a rank-change or rank-drop point of* G.

In the context of projector-based methods addressed in Chapter 2, critical points obstruct the construction of a tractability chain on the whole interval until the last step. On the other hand, the matrix chain helps us to characterize critical points. The following examples demonstrate how a critical point may arise in a linear DAE and how the matrix chain identifies such a critical point.

Example 3.3. Consider the DAE

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \Big(\begin{bmatrix} 1 & -t \end{bmatrix} x(t) \Big)' + \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} x(t) = 0, \quad t \in \mathcal{J} = (-\infty, \infty), \tag{3.1}$$

where the leading term is properly stated with $\ker A = \{0\}$, $\operatorname{im} D = \mathbb{R}$, $R = 1$. This DAE has a critical point at $t_* = 1$ since the matrix sequence can be formed only on the

intervals $(-\infty, 1)$ and $(1, \infty)$ but not on the whole interval $\mathcal{J}$. Setting

$$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & -t \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix},$$

and calculating the matrix chain we find

$$G_0 = AD = \begin{bmatrix} 1 & -t \\ 1 & -t \end{bmatrix}, \quad N_0 = \ker G_0 = \operatorname{span}\left\{ \begin{bmatrix} t \\ 1 \end{bmatrix} \right\}, \quad Q_0(t) = \begin{bmatrix} 0 & t \\ 0 & 1 \end{bmatrix},$$
$$G_1 = G_0 + B_0 Q_0 = \begin{bmatrix} 1 & 1 \\ 1 & 2-t \end{bmatrix},$$

with $\det G_1 = 1 - t$, $G_1^{-1} = \frac{1}{1-t} \begin{bmatrix} 2-t & -1 \\ -1 & 1 \end{bmatrix}$. Due to the rank deficiency at $t_* = 1$ of G_1, $t_* = 1$ is a critical point, whereas all points other than 1 are regular.

On the intervals $(-\infty, 1)$ and $(1, \infty)$ the DAE is regular with tractability index 1 and its solutions are given by

$$x(t) \;=\; \frac{1}{1-t} \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t),$$

where u satisfies the inherent singular ODE

$$u'(t) + \frac{2}{1-t} u(t) \;=\; 0. \tag{3.2}$$

The homogeneous ODE (3.2) has the solutions $u(t) = (t-1)^2 u(0)$. Hence, the solutions of the equation (3.1) are

$$x(t) \;=\; (1-t)u(0) \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad u(0) \in \mathbb{R},$$

which shows that all solutions vanish at $t_* = 1$. Solutions cannot be therefore uniquely specified by prescribing their values at $t_* = 1$.

Critical points of DAEs do not only occur in the last step of the chain construction. In more involved cases, they may also appear in an earlier step.

Example 3.4. Consider the Hessenberg DAE

$$x_1' + x_2 + t x_3 = q_1(t), \tag{3.3a}$$
$$x_2' - x_1 = q_2(t), \tag{3.3b}$$
$$x_1 = q_3(t). \tag{3.3c}$$

This system can be written in the properly stated form (2.17) using

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 & t \\ -1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \tag{3.4}$$

with $\ker A = 0$, $\operatorname{im} D = \mathbb{R}^2$, $R = I$, and $D^- = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$. The DAE (3.3) has a critical point at $t_* = 0$, since the transversality requirement for admissible projectors fails at this point. The matrix chain can be performed in the following way:

$$G_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},\ N_0 = \ker G_0 = \operatorname{span}\left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\},\ Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$G_1 = \begin{bmatrix} 1 & 0 & t \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},\ N_1 = \ker G_1 = \operatorname{span}\left\{ \begin{bmatrix} -t \\ 0 \\ 1 \end{bmatrix} \right\}.$$

The matrix G_1 has constant rank $r_1 = 2$ and the intersection $N_0(t) \cap N_1(t)$ becomes non-trivial when $t = 0$,i.e., $N_0(0) \cap N_1(0) = N_0(0) \neq \{0\}$.
On $\mathbb{R}^-$ and $\mathbb{R}^+$ we may choose the projector function

$$Q_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ -\frac{1}{t} & 0 & 0 \end{bmatrix}$$

onto $\ker G_1$ to fulfill the condition $Q_1 Q_0 = 0$. This yields, for $t \neq t_* = 0$, the nonsingular matrix

$$G_2 = G_1 + B_1 Q_1 = \begin{bmatrix} 1 & 0 & t \\ -1 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

The DAE (3.3) is therefore regular with tractability index 2 on $\mathbb{R}^-$ and $\mathbb{R}^+$.

Observe that the intersection $N_0(t) \cap N_1(t) = \{0\}$ and the chosen projector Q_1 is unbounded at the critical point. If we choose another projector function

$$\tilde{Q}_1 = \begin{bmatrix} 0 & 0 & -t \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which is defined at the critical point, then the admissibility condition $\tilde{Q}_1 Q_0 = 0$ is not satisfied. In addition, on the intervals where $t \neq 0$, the solution of the DAE (3.3) is given, with $x_2(t_0) \in \mathbb{R}$, by

$$\begin{aligned} x_1(t) &= q_3(t) \\ x_2(t) &= x_2(t_0) + \int_{t_0}^{t} (q_2(s) + q_3(s))ds \\ x_3(t) &= \frac{1}{t}(q_1(t) - q_3'(t) - x_2(t)). \end{aligned}$$

Clearly, this system cannot be solved for x_3 at the critical point.

The above examples motivate the taxonomy of critical points defined in the next section.

3.1 Classification of critical points

As said in Definition 2.15, the assumptions supporting the tractability index of linear properly stated DAEs base upon certain constant rank, transversality and smoothness requirements. Critical points may then be defined for the cases when those conditions are not satisfied. Nevertheless, it has been recently shown [69] that if the DAE coefficients are sufficiently smooth (that is, C^{m-1}), then the critical points corresponding to the failure of condition (c) in Definition 2.15, in Section 2, can be avoided as described in Proposition 3.7 below.

Definition 3.5. *Consider the DAE* (2.17) *with continuous coefficients. The point* $t_* \in \mathcal{J} - \mathcal{J}_{reg}$ *is said to be a critical point of*

(i) type 0 *if* G_0 *has a rank drop at* t_**;*

(ii) type k*-A,* $k \geq 1$*, if there exists a neighborhood* $\mathcal{I} \subseteq \mathcal{J}$ *of* t_* *where the DAE has admissible projector functions* $Q_0, \ldots, Q_{k-1}$*, but* G_k *has a rank drop at* t_* *for some (hence any) admissible projector functions* $Q_0, \ldots, Q_{k-1}$*;*

(iii) type k*-B,* $k \geq 1$*, if there exists a neighborhood* $\mathcal{I} \subseteq \mathcal{J}$ *of* t_* *where the DAE has admissible projector functions* $Q_0, \ldots, Q_{k-1}$ *and* G_k *has constant rank for some (hence any) admissible projector functions* $Q_0, \ldots, Q_{k-1}$*, but the intersection* $\widehat{N}_k(t_*) := N_k(t_*) \cap (N_0(t_*) + \cdots + N_{k-1}(t_*))$ *is non-trivial, for these (hence any other) projector functions and* G_k*.*

Critical points of type 0, defined by the rank deficiencies of G_0, may follow from rank deficiencies of A or D, or of both, and from the failure of the transversality property of $\ker A$ and $\operatorname{im} D$ in (2.18).

Level-k critical points will be referred either to type k-A or type k-B. In addition, a critical point of type k-A or k-B with arbitrary k will be called a critical point of type A or B, respectively.

In [69] it has been shown that the classification of critical points are independent of the (admissible) projectors. Furthermore, following the proof in [65], this taxonomy can be proved to be invariant under linear time-varying coordinate changes and premultiplication by a nonsingular, continuous matrix function.

Theorem 3.6. *The definitions of critical points of types* 0*,* k*-A, and* k*-B are independent of the (admissible) choice of projector functions.*

Additionally, if t_* *is a critical point of type* k*-A or* k*-B,* $1 \leq k \leq m$*, or of type* 0*, for the DAE* (2.17)*, then* t_* *is a critical point of the same type for the rescaled, transformed DAE*

$$\tilde{A}(t)(\tilde{D}(t)y(t))' + \tilde{B}(t)y(t) = L(t)q(t), \quad t \in \mathcal{J}, \tag{3.5}$$

with non-singular $L(t), K(t) \in C(\mathcal{J}, L(\mathbb{R}^m))$*,* $\tilde{A}(t) = L(t)A(t)$*,* $\tilde{D}(t) = D(t)K(t)$*,* $\tilde{B}(t) = L(t)B(t)K(t)$*.*

Proof of the first assertion is given in [69]. The second assertion follows from the construction of the projectors $\tilde{Q}_i = K^{-1}Q_iK$ for (3.5) in [65], which results in the identity $\tilde{G}_i = LG_iK$. The rank of G_i is therefore transferred to $\tilde{G}_i$ and type 0 and type-A critical points are hence invariant. In addition, $N_i = \ker \tilde{G}_i = K^{-1}N_i$, so that the loss of transversality in the N_i spaces defining type-B critical points is also transferred to $\tilde{N}_i$.

As introduced above, sufficient smoothness of the linear DAE coefficients ensures the existence of the required smoothness properties in Definition 2.15. This means that only critical points of types A and B will arise in linear DAEs with sufficiently smooth coefficients. This statement is confirmed by the following proposition.

Proposition 3.7. *Assume that the coefficients $A(t)$, $D(t)$, $B(t)$ in the DAE (2.17) are C^{m-1}. Then every critical point is of type k-A or k-B, with $1 \le k \le m$, or of type 0.*

Proof : Let both $A(t)$ and $D(t)$ have constant rank on $\mathcal{I}$, and $\ker A(t) \oplus \operatorname{im} D(t) = \mathbb{R}^n$ for all $t \in \mathcal{I}$. This implies that no type 0 critical points are met. Then we can take a Q_0 in C^{m-1}, so that $G_1 = G_0 + BQ_0$ is also in C^{m-1}.
If G_1 is singular and neither type 1-A nor type 1-B critical points are met, then we may choose a C^{m-1} projector function Q_1 satisfying the condition $Q_1Q_0 = 0$, so that $D\Pi_1D^-$ is in C^{m-1}. Hence $B_1 = BP_0 - G_1D^-(D\Pi_1D^-)'DP_0$ and so G_2 will be in the class C^{m-2}.
If no critical points of types A or B arise in subsequent levels, we can continue to construct an admissible sequence up to G_{m-1}, Q_{m-1} in C^1, so that continuous B_{m-1} and G_m can be found. Now, if G_m is singular and has constant rank, i.e., regular points and type m-A critical point are ruled out, then $N_0 + \cdots + N_{m-1} = \mathbb{R}^m$, $\widehat{N}_m = N_m \cap \mathbb{R}^m = N_m \neq \{0\}$ and a critical point of type m-B is met. □

In the light of Definition 3.5, $t_* = 1$ in Example 3.3 is a critical point of type 1-A arising in the last step of the chain construction, whereas $t_* = 0$ in Example 3.4 is a critical point of type 1-B. For a type 0 critical point we give the following example:

Example 3.8. Consider the DAE

$$x_1 = q_1(t), \tag{3.6a}$$
$$tx_1' + x_2 = q_2(t), \tag{3.6b}$$

where $t \in \mathcal{J} = (-\infty, \infty)$. This equation can be written as

$$\begin{bmatrix} 0 & 0 \\ t & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} x(t) \right)' + x(t) = q(t),$$

which has a properly stated leading term on the intervals $(-\infty, 0)$ and $(0, \infty)$, with

$$A = \begin{bmatrix} 0 & 0 \\ t & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = I, \quad R = D^- = D.$$

This DAE has a type 0 critical point at $t_* = 0$, since $A(t)$ has a rank drop at this point. Note that, as mentioned above, a type 0 critical point may follow from a rank deficiency of the leading coefficient $A(t)$.
On $\mathbb{R}^-$ and $\mathbb{R}^+$ the matrices A, D, and D^- lead to

$$G_0 = AD = \begin{bmatrix} 0 & 0 \\ t & 0 \end{bmatrix}, \quad N_0(t) = \ker G_0(t) = \operatorname{span}\left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}.$$

Choosing $Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ as a projector function onto $\ker G_0$ we obtain

$$G_1 = G_0 + B_0 Q_0 = \begin{bmatrix} 0 & 0 \\ t & 1 \end{bmatrix}, \quad N_1(t) = \ker G_1(t) = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ -t \end{bmatrix} \right\}.$$

The matrix G_1 is singular and has constant rank. Then, defining a projector function $Q_1(t) = \begin{bmatrix} 1 & 0 \\ -t & 0 \end{bmatrix}$ onto $\ker G_1$ and calculating $B_1 = B_0 P_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ yield

$$G_2 \;=\; G_1 + B_1 Q_1 \;=\; \begin{bmatrix} 1 & 0 \\ t & 1 \end{bmatrix},$$

with $G_2^{-1}(t) = \begin{bmatrix} 1 & 0 \\ -t & 1 \end{bmatrix}$. Therefore, on $\mathbb{R}^-$ and $\mathbb{R}^+$ the DAE is regular with tractability index 2 and its solution is given by

$$x_1(t) = q_1(t), \quad x_2(t) = q_2(t) - t q_1'(t).$$

Observe that the solution of this problem is also uniquely determined at the critical point $t_* = 0$. In particular, the homogeneous DAE (3.6) with $q = 0$ has only the trivial solution. Note also that the resulting matrix $G_2(t)$ is nonsingular at the critical point; such *harmless* critical points will be addressed in Subsection 3.2.2.

3.2 A-Critical chain

As stated above, critical points obstruct the construction of the matrix chain on the whole interval $\mathcal{J}$. Nevertheless, under Assumptions 1 and 2 below, one can adapt the projector-based method addressed in Chapter 2 to handle linear DAEs which possess critical points [69] by extending projector functions continuously to the whole interval. These assumptions make it possible to form a matrix function sequence $\{G_i\}$ for a linear DAE (2.17) with critical points. Furthermore, the behavior of problems with critical points can be described via a scalarly implicit inherent ODE as stated in Theorem 3.12 below. This implicit ODE generalizes the inherent explicit ODE (2.40) in Chapter 2. Keep in mind that the regular set $\mathcal{J}_{reg}$ is defined as the union of the regularity intervals.

Assumption 1. *The set $\mathcal{J}_{reg}$ of regular points is dense in $\mathcal{J}$.*

Assumption 2. *There exist projector functions $Q_0, \ldots, Q_{m-1}$ on $\mathcal{J}$ such that, for $0 \le i \le m-1$:*

(a) Q_i is continuous in the whole $\mathcal{J}$;

(b) $Q_i(t)$ is the projector function onto $\ker G_i(t)$ with $\operatorname{im} Q_i(t) = N_i(t)$ for all $t \in \mathcal{J}_{reg}$;

(c) $Q_i Q_j = 0$ for all $t \in \mathcal{J}$, $0 \le j < i$;

(d) $D\Pi_i D^-$ is continuously differentiable in $\mathcal{J}_{reg}$ and $(D\Pi_i D^-)'$, $D^-(D\Pi_i D^-)'D$ have continuous extensions on $\mathcal{J}$.

The condition (c) in Assumption 2 is equivalent to the property

$$N_0 + \cdots + N_{i-1} \subseteq \ker Q_i, \quad i \ge 1,$$

i.e., it is equivalent to the empty intersection

$$\widehat{N}_i(t) := N_i(t) \cap (N_0(t) + \cdots + N_{i-1}(t)) = \{0\}$$

in the second admissibility requirement in Definition 2.15. This means that the projector functions defined by Assumptions 1 and 2 coincide with the admissible projector functions on the regular intervals.

Proposition 3.9. *Assumptions 1 and 2 rule out type B critical points on $\mathcal{J}$.*

Proof : Let Assumptions 1 and 2 be given. Suppose that t_* is a type k-B critical point for some $k \ge 1$. By Definition 3.5, there exist admissible projector functions $Q_0, \ldots, Q_{k-1}$ and the matrix function G_k has constant rank in some neighborhood of t_*. Furthermore, $N_k = \ker G_k$ has constant dimension around t_*. We have to show that $\widehat{N}_k(t_*) = \{0\}$. From the conditions (b) and (c) in Assumption 2, it results that

$$\{N_0(t) + \cdots + N_{k-1}(t)\} \cap \operatorname{im} Q_k(t) = \{0\}$$

for all t in $\mathcal{J}_{reg}$. Thus, it is sufficient to prove that $\operatorname{im} Q_k(t_*) = N_k(t_*)$. Both spaces have the same dimension due to the continuity of Q_k and the constant dimension of N_k; on the other hand, from the vanishing of the continuous product $G_k Q_k$ in the dense set $\mathcal{J}_{reg}$, it follows that $G_k(t_*)Q_k(t_*) = 0$, so that $\operatorname{im} Q_k(t_*) \subseteq \ker G_k(t_*) = N_k(t_*)$. This verifies that type k-B critical points are excluded for $1 \le k \le m$ by Assumptions 1 and 2. □

According to Proposition 3.9, the matrix chain $\{G_i\}$ built under Assumptions 1 and 2 is called an A-critical chain. Further detailed study on critical points of type B can be found in [70, 76].

Proposition 3.10. *Under Assumptions 1 and 2, the DAE has the same characteristic values and, in particular, the same index μ in the whole $\mathcal{J}_{reg}$.*

Proof : The assumed continuity on Q_i in item (a) of Assumption 2 means that its nullspace and image have constant dimension. This implies a constant rank condition on Q_i in the regular intervals $\mathcal{J}_{reg}$ and hence on G_i, since Q_i is the projector onto $\ker G_i$ on $\mathcal{J}_{reg}$. This proves that the characteristic value $r_i = \operatorname{rank} G_i$ is uniform on $\mathcal{J}_{reg}$ and the nonsingular matrix G_i defining the index will be arrived at the same level on the regular set $\mathcal{J}_{reg}$. □

Remark 3.11. *From condition (b) in Assumption 2, it follows that the relation* $\operatorname{im} Q_i(t) = N_i(t) = \ker G_i(t)$ *for all* $t \in \mathcal{J}_{reg}$ *is satisfied, so that we have* $\operatorname{im} Q_i(t) = N_i(t) \subseteq \ker G_i(t)$ *on the regular set* $\mathcal{J}_{reg}$*. Further, the continuity of the projector* Q_i *on the whole* $\mathcal{J}$ *assumed in item (a) of Assumption 2 implies that the nullspace and image of* Q_i *have constant dimension on* $\mathcal{J}$*. Then, if the matrix* G_i *has a rank drop at* t_**, but* Q_i *has constant rank, then* $\operatorname{im} Q_i(t_*)$ *is strictly contained in* $\ker G_i(t_*)$*, that is,* $\operatorname{im} Q_i(t_*) \subset \ker G_i(t_*)$*. Consequently, under Assumptions 1 and 2, the property* $\operatorname{im} Q_i(t) = N_i(t) \subseteq \ker G_i(t)$ *is satisfied on the whole* $\mathcal{J}$*. This property motivates the construction of a quasi-proper matrix chain as addressed in the next chapter.*

3.2.1 Decoupling for DAEs with critical points

Under Assumptions 1 and 2 the dynamical behavior of linear DAEs with critical points can be described in terms of an *implicit inherent* ODE. The leading coefficient of the scalarly implicit ODE does not vanish at regular points. That means on regular subintervals this implicit inherent ODE coincides with an inherent explicit regular ODE. The following result is taken from [69].

Theorem 3.12. *Under Assumptions 1 and 2,*

$$x \in C_D^1(\mathcal{I}, \mathbb{R}^m) \;:=\; \{x \in C(\mathcal{I}, \mathbb{R}^m) : (Dx) \in C^1(\mathcal{I}, \mathbb{R}^n)\}$$

is a solution of the DAE (2.17) *in a subinterval* $\mathcal{I} \subseteq \mathcal{J}$ *if and only if it can be written as*

$$x \;=\; D^- u + v_0 + \cdots + v_{\mu-1}, \tag{3.7}$$

where $u \in C^1(\mathcal{I}, \mathbb{R}^n)$ *is a solution of the* scalarly implicit inherent *ODE*

$$\omega_\mu u' - \omega_\mu (D\Pi_{\mu-1}D^-)' u + D\Pi_{\mu-1} G_\mu^{adj} B D^- u \;=\; D\Pi_{\mu-1} G_\mu^{adj} q, \tag{3.8}$$

lying on the locally invariant space $\operatorname{im} D\Pi_{\mu-1}$*, whereas the components* $v_0 \in C(\mathcal{I}, \mathbb{R}^m)$ *and* $v_i \in C_D^1(\mathcal{I}, \mathbb{R}^m)$*,* $i = 1, \ldots, \mu-1$*, satisfy*

$$\omega_\mu v_{\mu-1} \;=\; \mathcal{L}_{\mu-1}^{adj} q - \mathcal{K}_{\mu-1}^{adj} D^- u, \tag{3.9a}$$

$$\begin{aligned} \omega_\mu v_i \;=\; & \mathcal{L}_i^{adj} q - \mathcal{K}_i^{adj} D^- u + \omega_\mu \sum_{j=i+1}^{\mu-1} \mathcal{N}_{ij} (Dv_j)' \\ & + \omega_\mu \sum_{j=i+2}^{\mu-1} \mathcal{M}_{ij} v_j, \quad i = 0, \ldots, \mu-2. \end{aligned} \tag{3.9b}$$

Here, the scalar function $\omega_\mu(t)$ stands for $\det G_\mu(t)$, G_μ^{adj} is the adjoint of G_μ, namely the transpose of the matrix of cofactors of G_μ, and μ means the index of the DAE (2.17) on $\mathcal{J}_{reg}$. The coefficients $\mathcal{L}_{\mu-1}^{adj}$ and $\mathcal{K}_{\mu-1}^{adj}$ in (3.9a) read

$$\begin{aligned}\mathcal{L}_{\mu-1}^{adj} &= \Pi_{\mu-2}Q_{\mu-1}G_\mu^{adj}, \quad \mathcal{K}_{\mu-1}^{adj} = \Pi_{\mu-2}Q_{\mu-1}G_\mu^{adj}B\Pi_{\mu-1} \text{ for } \mu \geq 2,\\ \mathcal{L}_{\mu-1}^{adj} &= Q_{\mu-1}G_\mu^{adj}, \quad \mathcal{K}_{\mu-1}^{adj} = Q_{\mu-1}G_\mu^{adj}B\Pi_{\mu-1} \text{ for } \mu = 1.\end{aligned}$$

For $i = 1, \ldots, \mu-1$, $j = i+2, \ldots, \mu-1$, the coefficients $\mathcal{L}_i^{adj}$, $\mathcal{K}_i^{adj}$, $\mathcal{N}_{ij}$, $\mathcal{M}_{ij}$ are given by

$$\begin{aligned}\mathcal{L}_0^{adj} &= Q_0P_1\cdots P_{\mu-1}G_\mu^{adj}\\ \mathcal{L}_i^{adj} &= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{adj}, \quad i = 1, \ldots, \mu-1,\\ \mathcal{N}_{01} &= Q_0Q_1D^-, \quad \mathcal{N}_{0j} = Q_0P_1\cdots P_{j-1}Q_jD^-, \quad j = i+2, \ldots, \mu-1,\\ \mathcal{N}_{i,i+1} &= \Pi_{i-1}Q_iQ_{i+1}D^-, \quad \mathcal{N}_{ij} = \Pi_{i-1}Q_iP_{i+1}\cdots P_{j-1}Q_jD^-,\\ \mathcal{K}_0^{adj} &= Q_0P_1\cdots P_{\mu-1}G_\mu^{adj}B\Pi_{\mu-1}+\\ &\quad + \omega_\mu Q_0P_1\cdots P_{\mu-1}P_0D^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}\\ \mathcal{K}_i^{adj} &= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{adj}B\Pi_{\mu-1}+\\ &\quad + \omega_\mu\Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}P_iD^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}\\ \mathcal{M}_{ij} &= \Pi_{i-1}Q_i\{Q_{i+1}D^-(D\Pi_iQ_{i+1}D^-)' + P_{i+1}Q_{i+2}D^-(D\Pi_{i+1}Q_{i+2}D^-)'\\ &\quad + \cdots + P_{i+1}\cdots P_{\mu-2}Q_{\mu-1}D^-(D\Pi_{\mu-2}Q_{\mu-1}D^-)'\}D\Pi_{j-1}Q_j\\ &\quad - \sum_{l=1}^{j}\Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}\cdots P_lD^-(D\Pi_lD^-)'D\Pi_{j-1}Q_j.\end{aligned}$$

Proof of this statement can be found in [69, 76]. The idea of the proof relies upon the fact that the decoupling procedure in Theorem 2.23 holds on the dense subset $\mathcal{I} \cap \mathcal{J}_{reg}$, and remains valid also on the whole interval $\mathcal{I}$. Obviously, via the identity $\det G_\mu(t) \cdot I = G_\mu^{adj}(t) \cdot G_\mu(t)$, on the intervals where $\omega_\mu(t) = \det G_\mu(t) \neq 0$ (or in regular intervals) the system (3.8)–(3.9) coincides with the decoupling obtained with the classical projector-based methods in Section 2.

3.2.2 Harmless critical points

An exemplary introduction to harmless critical points has been given in Example 3.8 in the previous section. In this section we introduce harmless critical points formally. As detailed above, Assumptions 1 and 2 allow us to adapt the projector-based method to a linear DAE with critical points and make it possible to characterize the solutions of such a DAE through an implicit inherent ODE as depicted in Theorem 3.12. The leading coefficient $\omega_\mu(t) = \det G_\mu(t)$ within (3.8)–(3.9) will vanish if and only if $G_\mu(t)$ is singular, where μ is the index of the DAE (2.17) on the regular intervals. However, this is not always the case as illustrated in Example 3.8. There, although $t_* = 0$ is the critical point, the resulting matrix $G_2(t)$ remains

nonsingular at this point.

Definition 3.13. *Under Assumptions 1 and 2, a critical point t_* is said to be harmless, if the resulting matrix function $G_\mu(t)$ is nonsingular at t_*, where μ is the index of the DAE on the interval $\mathcal{J}_{reg}$.*

Proposition 3.14. *Under Assumptions 1 and 2, a type $(\mu-1)$-A critical point t_* leads to a singular $G_\mu(t_*)$.*

Proof : Let Assumptions 1 and 2 be given and t_* be a type $(\mu-1)$-A critical point. Then $G_{\mu-1}$ has a rank drop at t_*, but $Q_{\mu-1}$ has constant rank by Assumption 2-(a). Thus $\operatorname{im} Q_{\mu-1}(t_*) \subset \ker G_{\mu-1}(t_*)$, that is, there exists a nontrivial vector $z \in \ker G_{\mu-1}(t_*) \setminus \operatorname{im} Q_{\mu-1}(t_*)$. This implies

$$G_{\mu-1}(t_*)z \;=\; 0 \quad \text{and} \quad P_{\mu-1}(t_*)z \;=\; (I - Q_{\mu-1}(t_*))z \;\neq\; 0.$$

Then, computing

$$\begin{aligned} G_\mu(t_*)P_{\mu-1}(t_*)z &= (G_{\mu-1}(t_*) + B_{\mu-1}(t_*)Q_{\mu-1}(t_*))P_{\mu-1}(t_*)z \\ &= G_{\mu-1}(t_*)P_{\mu-1}(t_*)z \\ &= G_{\mu-1}(t_*)z - G_{\mu-1}(t_*)Q_{\mu-1}(t_*))z \\ &= 0, \end{aligned}$$

yields $P_{\mu-1}(t_*)z \in \ker G_\mu(t_*)$ and proves that $G_\mu(t_*)$ is singular. □

From Proposition 3.14 follows the next result.

Corollary 3.15. *Under Assumptions 1 and 2, a necessary condition for a critical point t_* to be harmless is that $G_{\mu-1}$ has constant rank in some neighborhood of t_*.*

The following example demonstrates harmless critical points for index-2 DAE on $\mathcal{J}_{reg}$ and the corresponding regular inherent ODEs:

Example 3.16. Consider the DAE

$$\begin{array}{lcllcllcll} \alpha(t)x_3'(t) & + & x_2(t) & & & & & = & q_1(t), & \\ \beta(t)x_2'(t) & - & x_3(t) & + & x_4(t) & & & = & q_2(t), & \\ \gamma(t)x_1'(t) & + & x_3(t) & + & x_5(t) & & & = & q_3(t), & (3.10) \\ & & x_2(t) & + & x_4(t) & + & x_5(t) & = & q_4(t), & \\ & & x_1(t) & + & x_3(t) & & & = & q_5(t), & \end{array}$$

where $\alpha(t)$, $\beta(t)$, and $\gamma(t)$ are continuous differentiable functions, $\beta(t) \neq 0$, and $t \in \mathcal{J} = \mathbb{R}$. This system can be written in the formulation (2.17) using

$$A(t) = \begin{bmatrix} \alpha(t) & 0 & 0 \\ 0 & \beta(t) & 0 \\ 0 & 0 & \gamma(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The matrix function $A(t)$ has rank-drop points at the zeros of $\alpha(t)$, $\beta(t)$, and $\gamma(t)$. The product $G_0(t) = A(t)D$ read

$$G_0(t) \;=\; \begin{bmatrix} 0 & 0 & \alpha(t) & 0 & 0 \\ 0 & \beta(t) & 0 & 0 & 0 \\ \gamma(t) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and the nullspace $N_0(t)$ is defined by

$$N_0(t) \;=\; \ker G_0(t) \;=\; \{z \in \mathbb{R}^5 : \gamma(t)z_1 = 0, \beta(t)z_2 = 0, \alpha(t)z_3 = 0\}.$$

The function $\beta(t)$ has no zeros. The zeros of $\alpha(t)$ or $\gamma(t)$ yield rank deficiencies in the matrix functions $A(t)$ and $G_0(t)$, and hence define type 0 critical points of the problem.

Denote

$$\check{\mathcal{J}}_{reg} \;:=\; \{t \in \mathcal{J} : \alpha(t) \neq 0, \beta(t) \neq 0, \gamma(t) \neq 0\}$$

and we will show that $\check{\mathcal{J}}_{reg}$ is the set of regular points.

On $\check{\mathcal{J}}_{reg}$ the leading term of DAE (3.10) is properly stated and we may take

$$D^- = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad R = DD^- = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

to obtain continuous projector functions $Q_0 = \mathrm{diag}(0, 0, 0, 1, 1)$ and $\Pi_0 = P_0 = I - Q_0 = \mathrm{diag}(1, 1, 1, 0, 0)$ in the whole $\mathcal{J}$, i.e., Assumption 2-(a) is met. Note that Q_0 is the projector onto $\ker G_0$ on $\check{\mathcal{J}}_{reg}$ and that Assumption 2-(b) holds. In addition, the properly stated property specified in Definition 2.14 implies the condition in Assumption 2-(d). Hence, Assumption 2 is satisfied for the first level.

Then, we have

$$G_1(t) \;=\; G_0(t) + BQ_0 \;=\; \begin{bmatrix} 0 & 0 & \alpha(t) & 0 & 0 \\ 0 & \beta(t) & 0 & 1 & 0 \\ \gamma(t) & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The matrix $G_1(t)$ has constant rank $r_1 = 4$ on $\check{\mathcal{J}}_{reg}$ and the nullspace $N_1(t) = \ker G_1(t)$ reads

$$N_1(t) \;=\; \{z \in \mathbb{R}^5 : z_2 = -\frac{\gamma(t)}{\beta(t)}z_1, z_3 = 0, z_4 = \gamma(t)z_1, z_5 = -\gamma(t)z_1\}.$$

For $t \in \check{\mathcal{J}}_{reg}$, we choose the continuous projector function $Q_1(t)$ onto $\ker G_1(t)$ to fulfill

Assumption 2-(b) as

$$Q_1(t) = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ -\frac{\gamma(t)}{\beta(t)} & 0 & -\frac{\gamma(t)}{\beta(t)} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \gamma(t) & 0 & \gamma(t) & 0 & 0 \\ -\gamma(t) & 0 & -\gamma(t) & 0 & 0 \end{bmatrix},$$

which is continuous on the whole $\mathcal{J}$, meeting Assumption 2-(a). The projector Q_1 verifies the condition $Q_1Q_0 = 0$ on $\mathcal{J}$, so that Assumption 2-(c) holds. Calculating

$$\Pi_1(t) = P_0P_1 = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ \frac{\gamma(t)}{\beta(t)} & 1 & \frac{\gamma(t)}{\beta(t)} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad D\Pi_1(t)D^- = \begin{bmatrix} 1 & 0 & 0 \\ \frac{\gamma(t)}{\beta(t)} & 1 & \frac{\gamma(t)}{\beta(t)} \\ -1 & 0 & 0 \end{bmatrix},$$

leads to $B_1(t)$ and $G_2(t) = G_1(t) + B_1(t)Q_1(t)$ on $\check{\mathcal{J}}_{reg}$ as

$$B_1(t) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ -\beta(t)\left(\frac{\gamma(t)}{\beta(t)}\right)' & 0 & -\beta(t)\left(\frac{\gamma(t)}{\beta(t)}\right)' - 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix},$$

$$G_2(t) = \begin{bmatrix} -\frac{\gamma(t)}{\beta(t)} & 0 & \alpha(t) - \frac{\gamma(t)}{\beta(t)} & 0 & 0 \\ -\beta(t)\left(\frac{\gamma(t)}{\beta(t)}\right)' & \beta(t) & -\beta(t)\left(\frac{\gamma(t)}{\beta(t)}\right)' & 1 & 0 \\ \gamma(t) & 0 & 0 & 0 & 1 \\ -\frac{\gamma(t)}{\beta(t)} & 0 & -\frac{\gamma(t)}{\beta(t)} & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Obviously, Assumption 2 is satisfied at this level. Due to $\det G_2(t) = \alpha(t)\beta(t)$, the resulting matrix $G_2(t)$ is nonsingular for $t \in \check{\mathcal{J}}_{reg}$. Therefore, the DAE (3.10) is regular with tractability index 2. It should be mentioned that Assumption 1 holds because the set of regular points $\check{\mathcal{J}}_{reg}$ is dense in $\mathcal{J}$. In addition, the zeros of $\gamma(t)$ define harmless critical points, since the matrix $G_2(t)$ is also nonsingular at those points.

Now, we represent how the functions $\alpha(t)$, $\beta(t)$, and $\gamma(t)$ affect the inherent explicit regular ODE which arises in the decoupling procedure discussed in Theorem 3.12. In the following, we omit the argument t for simplicity. From the above A-critical chain, the canonical projector Π_{can2} is defined by

$$\begin{aligned} \Pi_{can2} &:= \left(I - Q_0P_1G_2^{-1}B\Pi_1 + Q_0Q_1D^-\left(D\Pi_1D^-\right)'D\Pi_1\right)\Pi_1 \\ &= \frac{1}{\alpha\beta}\begin{bmatrix} 0 & 0 & -\alpha\beta & 0 & 0 \\ \alpha\gamma & \alpha\beta & \alpha\gamma & 0 & 0 \\ 0 & 0 & \alpha\beta & 0 & 0 \\ -\gamma(\alpha-\gamma) & -\beta(\alpha-\gamma) & \alpha\beta - \gamma(\alpha-\gamma) & 0 & 0 \\ -\gamma^2 & -\beta\gamma & -\alpha\beta - \gamma^2 & 0 & 0 \end{bmatrix}, \end{aligned}$$

where

$$G_2^{-1} = \frac{1}{\alpha\beta}\begin{bmatrix} -\beta & 0 & 0 & 0 & \alpha\beta-\gamma \\ \gamma & \alpha & \alpha & -\alpha & \alpha\beta\left(\frac{\gamma}{\beta}\right)' - \alpha\gamma(\omega+1) \\ \beta & 0 & 0 & 0 & \gamma \\ -\beta\gamma & 0 & -\alpha\beta & \alpha\beta & \alpha\beta\gamma(\omega+1) \\ \beta\gamma & 0 & \alpha\beta & 0 & -\gamma(\alpha\beta-\gamma) \end{bmatrix}.$$

Then, on $\check{\mathcal{J}}_{reg}$, solutions of the DAE (3.10) are given by

$$\begin{aligned} x &= D^- u + v_0 + v_1 \\ &= \begin{bmatrix} u_3 \\ u_2 \\ u_1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -\frac{\gamma}{\alpha}q_1 - q_3 + q_4 + \gamma\omega q_5 + \gamma q_5' + (1-\gamma\omega)u_1 - \beta\omega u_2 - \gamma\omega u_3 \\ \frac{\gamma}{\alpha}q_1 + q_3 + \frac{\gamma^2}{\alpha\beta}q_5 - \gamma q_5' - (1+\frac{\gamma^2}{\alpha\beta})u_1 - \frac{\gamma}{\alpha}u_2 - \frac{\gamma^2}{\alpha\beta}u_3 \end{bmatrix} \\ &\quad + \frac{1}{\alpha\beta}\begin{bmatrix} \alpha\beta q_5 \\ -\alpha\gamma q_5 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \end{aligned}$$

where $\omega(t) := \frac{\alpha(t)-\gamma(t)}{\alpha(t)\beta(t)}$ and $u(t) = D\Pi_1(t)x(t)$ solves the regular ODE of the form

$$u' + \frac{1}{\alpha\beta}\begin{bmatrix} \gamma & \beta & \gamma \\ -\alpha\gamma\omega & -\alpha+\gamma & -\alpha\gamma\omega \\ -\gamma & -\beta & -\gamma \end{bmatrix} u = \frac{1}{\alpha\beta}\tilde{q}, \tag{3.11}$$

with

$$\tilde{q} := \begin{bmatrix} \beta q_1 + \gamma q_5 \\ \gamma q_1 + \alpha(q_2 + q_3 - q_4) + \alpha\left(\beta\left(\frac{\gamma}{\beta}\right)' - \gamma\omega\right) q_5 \\ -\beta q_1 - \gamma q_5 \end{bmatrix}.$$

The solutions of the homogeneous ODE, with $q_1 = 0$, $q_4 = q_2 + q_3$, $q_5 = 0$, are

$$\begin{aligned} u_1(t) &= K_2 - u_3(t) \\ u_2(t) &= e^{\int_{t_0}^t \omega(s)ds}\left(K_1 + K_2\int_{t_0}^t e^{-\int_{t_0}^s \omega(r)dr}\frac{(\alpha(s)-1)\beta^2(s)\gamma(s)}{\alpha(s)}ds\right) \\ u_3(t) &= K_3 + \int_{t_0}^t (K_2\gamma(s) + \beta(s)u_2(s))\frac{1}{\alpha(s)\beta(s)}ds \end{aligned}$$

for K_1, K_2, $K_3 \in \mathbb{R}$.

Observe that the solutions of the homogeneous ODE and hence the solutions of the DAE are well defined also at harmless critical points, that is, at the zeros of $\gamma(t)$. Furthermore, the resulting explicit regular ODE (3.11) will become a singular one at the points where $\alpha(t)$ and $\beta(t)$ vanish identically. Therefore, the zeros of $\alpha(t)$ and $\beta(t)$ can never define harmless critical points.

Chapter 4

Quasi-Regular Linear DAEs

As stated earlier, if the matrix function G_i changes the rank, its nullspace is discontinuous and a critical point of linear DAEs (2.17) exists (cf. Definition 3.5). Assuming the existence of continuous extensions of projector functions from the regularity set to the entire interval, harmless critical points can be characterized as stated in Chapter 3. However, these assumptions restrict the DAE to have the same index μ in the whole regular subset. Following [54], this limitation can be avoided by choosing a continuous subnullspace N_i belonging to the kernel of G_i instead of the discontinuous nullspace N_i. This allows the index of the DAE to change its value on the whole interval.

The following example show the advantage of replacing the discontinuous nullspace N_i by a continuous subnullspace N_i of $\ker G_i$.

Example 4.1. The DAE

$$\alpha(t)x_2'(t) + x_1(t) = q_1(t), \tag{4.1a}$$
$$x_2(t) = q_2(t), \tag{4.1b}$$

with scalar continuous function α on the interval $\mathcal{I} = [-1, 1]$,

$$\alpha(t) = 0 \quad \text{for} \quad t \leq 0, \quad \alpha(t) \neq 0 \quad \text{for} \quad t > 0,$$

has exactly one solution

$$x_2(t) = q_2(t), \quad t \in [-1, 1]$$
$$x_1(t) = \begin{cases} q_1(t), & t \in [-1, 0], \\ q_1(t) - \alpha(t)q_2'(t), & t \in (0, 1], \end{cases}.$$

This system can be written in the form (2.17) as

$$\begin{bmatrix} 0 & \alpha(t) \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 \\ 0 & \mathbb{1}_\alpha(t) \end{bmatrix} x \right)' + x = q, \tag{4.2}$$

with the characteristic function $\mathbb{1}_\alpha$ of the function α. However, this coefficient D is discontinous at $t_* = 0$ and $G_0 = AD$ undergoes rank deficiency at this point. The DAE is regular with index $\mu = 1$ and characteristic values $r_0 = 0$, $r_1 = 2$ on $[-1, 0]$ and regular with index $\mu = 2$ and characteristic values $r_0 = 1$, $r_1 = 1$, $r_2 = 2$ on $(0, 1]$.

Observe that, in contrast to Example 3.8, the matrix G_0 has different rank on subintervals. According to Proposition 3.10 the nullspace projector function does not have a continuous extension on the entire $\mathcal{J}$. Therefore, the Assumptions 1 and 2 do not hold for this system.
Let us consider the solution of this system. The solvability statements for regular DAEs show that functions q_2 are continuous on $[-1,0]$ and continuously differentiable on $(0,1]$. For instance, if

$$\alpha(t) = \begin{cases} 0\ , & t \in [-1,0], \\ t^{\frac{1}{3}}, & t \in (0,1], \end{cases}$$

$q_1(t) = 0$ and $q_2(t) = \alpha(t)$, we obtain

$$x_2(t) = \alpha(t) \quad \text{and} \quad x_1(t) = \begin{cases} 0, & t \in [-1,0], \\ -\frac{1}{3}t^{-\frac{1}{3}}, & t \in (0,1]. \end{cases}$$

As shown in Figure 4.1, the solution segments of the second component can be glued together smoothly, whereas this is not possible for the first component.

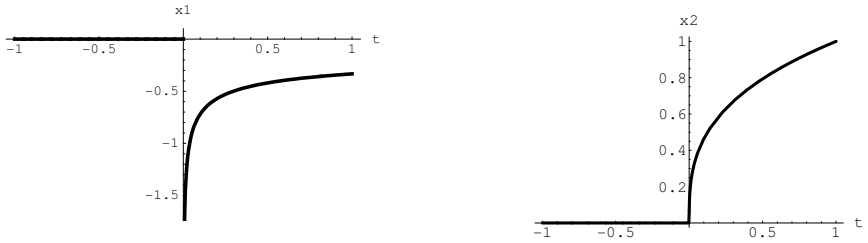


Figure 4.1: The solutions x_1, x_2 in case of $q_1 = 0$ and $q_2 = \alpha$

If we relax the strong solvability concept and choose smoother functions q, a continuous solution may be available on the whole interval $\mathcal{I}$. For example, for $q_1(t) = 0$, $q_2(t) = t^2$, the particular solution is $x_2(t) = t^2$ and $x_1(t) = \begin{cases} 0\ , & t \in [-1,0], \\ -2t^{\frac{4}{3}}, & t \in (0,1]. \end{cases}$ See Figure 4.2.

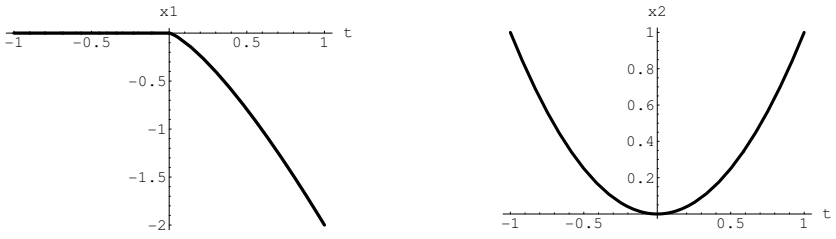


Figure 4.2: The solutions x_1, x_2 in case of smoother function q

More precisely, if we rewrite the DAE (4.1) as

$$\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} x \right)' + x = q, \tag{4.3}$$

we seek solutions that are continuous on the entire interval $\mathcal{I}$ with continuously differentiable component x_2, but now only continuous right-hand sides q with a continuously differentiable component q_2 (on [-1,1]) are considered. Further, the coefficient D is continuous on the whole of $\mathcal{I}$. Note that $R = D$ is a projector function and $\ker R = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}$ is a C^1-subspace of the kernel of A on the interval $\mathcal{I}$. The DAE (4.3) has a quasi-proper leading term, see below, on the interval $\mathcal{I}$.

4.1 Quasi-proper leading terms

The main idea to define a quasi-regular DAE [54] is to use a continuous subnullspace of the matrix G_i instead of the discontinuous nullspace $N_i := \ker G_i$.

Definition 4.2. *The leading term of the DAE* (2.17) *is said to be* quasi-properly stated *on the interval $\mathcal{I} \subseteq \mathcal{J}$ if the coefficient D has a nontrivial nullspace,* $\operatorname{im} D$ *is a C^1-subspace, and there exists a projector function $R \in C^1(\mathcal{I}, \mathbb{R}^n)$ such that* $\ker R(t) \subseteq \ker A(t)$*, $t \in \mathcal{I}$, and the transversality condition*

$$\operatorname{im} R(t) = \operatorname{im} D(t), \quad \ker R(t) \oplus \operatorname{im} D(t) = \mathbb{R}^n, \qquad t \in \mathcal{I}, \tag{4.4}$$

is valid.

Due to the C^1 structure on $\operatorname{im} D$, the continuous matrix function $D(t)$ has constant rank r and its nullspace $\ker D$ is a nontrivial continuous subspace with dimension $m-r$. The constant rank assumption on $D(t)$ implies that $\ker D(t)$ has a continuous basis on $\mathcal{I}$. This yields the existence of continuous projectors P_0, Q_0, and hence of a continuous generalized inverse D^- of D satisfying the relations

$$DD^-D = D, \quad D^-DD^- = D^-, \quad DD^- = R, \quad D^-D = P_0. \tag{4.5}$$

Remark 4.3. *At regular points where the leading term in* (2.17) *is (locally) properly stated, the subspace* $\ker R$ *coincides with* $\ker A$*. In general, in a quasi-proper leading term, the matrix function A has a C^1 time-varying subnullspace intersecting transversally with* $\operatorname{im} D$ *(see* [54]*).*

4.2 Quasi-regularity

In an analogous way as we did for regular DAEs, we build a sequence of matrix functions and subspaces for the DAE (2.17) in order to characterize a DAE with a quasi-proper leading term [54]. As previously, we omit the argument t for notional simplicity. All these definitions are meant pointwise for $t \in \mathcal{J}$. Assume that the leading term of (2.17) is quasi-proper on $\mathcal{I} \subseteq \mathcal{J}$. Set

$$G_0 := AD, \quad B_0 := B, \quad N_0 := \ker D \subseteq \ker AD = \ker G_0. \tag{4.6}$$

Since the matrix D has constant rank, we can choose the projector functions P_0, $Q_0 \in C(\mathcal{I}, L(\mathbb{R}^m))$ such that

$$\operatorname{im} Q_0 = N_0 := \ker D, \quad P_0 := I - Q_0, \quad \Pi_0 = P_0, \tag{4.7}$$

and define a continuous generalized inverse D^- fulfilling the properties (4.5) above. For $i \geq 0$, we compute, as long as the expressions exist,

$$G_{i+1} \;=\; G_i + B_i Q_i, \tag{4.8}$$

choose a continuous subspace N_{i+1},

$$N_{i+1} \;\subseteq\; \ker G_{i+1}, \tag{4.9}$$

projector functions Q_{i+1}, P_{i+1} such that

$$\operatorname{im} Q_{i+1} = N_{i+1}, \quad P_{i+1} := I - Q_{i+1},$$

and define

$$\begin{aligned} \Pi_{i+1} &= \Pi_i P_{i+1}, \\ B_{i+1} &= B_i P_i - G_{i+1} D^- (D\Pi_{i+1} D^-)' D\Pi_i. \end{aligned} \tag{4.10}$$

Denote $r_i := \operatorname{rank} G_i$, $i \geq 0$. The symbol Π_i is used, as previously, to stand for the product $P_0 \cdots P_i$. For the matrix function sequence (4.6)–(4.10) the identities $G_i z = 0$, $Q_i z = 0$ drive $G_{i+1} z = G_i z + B_i Q_i z = 0$, $i \geq 0$. Thus, the relation

$$(\ker G_i) \cap (\operatorname{im} P_i) \;\subseteq\; \ker G_{i+1}, \quad i \geq 0, \tag{4.11}$$

is satisfied.

It is important to notice that the matrix chain construction (4.6)–(4.10) is quite similar to the one defined for regular DAEs in Chapter 2. The only difference is that here the subspace N_{i+1} may be a subnullspace of the matrix G_{i+1}, whereas N_{i+1} in Chapter 2 must equal $\ker G_{i+1}$.

As in Chapter 2, we restrict the variety of possible projector functions further.

Definition 4.4. *Let* (2.17) *be a DAE with quasi-proper leading term on* $\mathcal{I} \subseteq \mathcal{J}$. *Any continuous projector function* Q_0 *onto* $\ker D$ *is said to be quasi-admissible for this DAE. The projector functions* $Q_0, \ldots, Q_k$ *with* $k \in \mathbb{N}$, *are said to be quasi-admissible on* $\mathcal{I}$ *for the DAE* (2.17) *if the following properties hold:*

(a) Q_i, $i = 1, \ldots, k$, *is continuous and* $N_i \subseteq \ker G_i$ *satisfies the condition*

$$\widehat{N}_i \;:=\; N_i \cap (N_0 + \cdots + N_{i-1}) \;=\; \{0\}, \tag{4.12}$$

$$N_0 + \cdots + N_{i-1} \;\subseteq\; \ker Q_i. \tag{4.13}$$

(b) Π_i *is continuous on* $\mathcal{I}$ *and* $D\Pi_i D^-$ *is continuously differentiable,* $i = 0, \ldots, k$.

Due to the trivial intersections $\widehat{N}_i = \{0\}$, $i = 1, \ldots, k$, it is possible to choose the continuous projector Q_i onto $N_i \subseteq \ker G_i$, $i = 1, \ldots, k$, in a way such that

$$N_0 + \cdots + N_{i-1} \subseteq \ker Q_i, \quad i = 1, \ldots, k. \tag{4.14}$$

This implies that the property

$$Q_i Q_j = 0, \quad 0 \leq j < i, \quad i = 1, \ldots, k,$$

is also fulfilled in the quasi-admissible framework.

Definition 4.5. *A DAE* (2.17) *with quasi-proper leading term is said to be* quasi-regular *on $\mathcal{I} \subseteq \mathcal{J}$ if there are an integer k and quasi-admissible projectors Q_0, …, Q_{k-1} such that the matrix G_k is nonsingular.*

Definition 4.6. *A DAE* (2.17) *with quasi-proper leading term is said to be* quasi-regular with index 1 *on $\mathcal{I} \subseteq \mathcal{J}$ if the DAE is quasi-regular on $\mathcal{I}$ with the nonsingular matrix G_2 and the projector product $Q_0 Q_1 = 0$.*

Since the subspace N_{k-1} is just a part of the nullspace $\ker G_{k-1}$, we can decompose

$$\ker G_{k-1} = N_{k-1} \oplus (\ker G_{k-1} \cap \operatorname{im} P_{k-1}).$$

Due to the non-singularity of G_k, the relation (4.11) reads, for $i = k - 1$,

$$\ker G_{k-1} \cap \operatorname{im} P_{k-1} = 0,$$

which implies the relation $N_{k-1} = \ker G_{k-1}$. Consequently, the constant rank property of G_{k-1} follows from the constant dimension of N_{k-1}. This makes it possible to extend the definition of harmless critical point in case of the quasi-regular DAEs.

Definition 4.7. *Given the DAE with quasi-proper leading term on $\mathcal{J}$, a critical point $t_* \in \mathcal{J}$ in the sense of Definition 2.18 is said to be harmless if there exists an open interval $\mathcal{I} \subseteq \mathcal{J}$ of t_* where the DAE is quasi-regular.*

Example 4.8. We consider again the DAE (4.1) formulated with quasi-proper leading term as (4.3) from Example 4.1. Here we choose a quasi-admissible projector Q_0 onto $\ker D$ and $\Pi_0 = I - Q_0$ as

$$Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Pi_0 = P_0 = R = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The matrix G_1 and the nullspace N_1 read

$$G_1 = \begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix}, \quad N_1 = \{z \in \mathbb{R}^2 : z_1 + \alpha z_2 = 0\}.$$

Taking

$$Q_1 = \begin{bmatrix} 0 & -\alpha \\ 0 & 1 \end{bmatrix}, \quad P_1 = \begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix},$$

we find

$$P_0P_1 = \Pi_1 = 0,\ \ Q_0Q_1 = \begin{bmatrix} 0 & -\alpha \\ 0 & 0 \end{bmatrix},\ \ B_1 = BP_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},\ \ G_2 = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix}.$$

Then, we have $r_2 = 2$, $r_1 = 1$, but $r_0(t) = 0$ if $t \le 0$, $r_0(t) = 1$ if $t > 0$. Observe that $N_0(t)$ is a proper subspace of $\ker G_0(t) = \mathbb{R}^2$, but $N_1(t)$ coincides with $\ker G_1(t)$ on $[-1, 1]$. Further, the critical point $t_* = 0$ is harmless (cf. Definition (4.7)), since the resulting matrix G_2 is nonsingular there.

Note that if the continuous subspaces N_i coincide with the nullspaces $\ker G_i$ almost everywhere on $\mathcal{J}$, then the projector functions are actually continuous extensions of the related nullspace projector functions, and the rank functions r_i are almost constant on $\mathcal{J}$. This situation was investigated by means of smooth projector extensions as addressed in Chapter 3.

4.3 Decoupling of quasi-regular DAEs

Let (2.17) be a quasi-regular DAE on the given interval $\mathcal{I} \subseteq \mathcal{J}$ that have a quasi-proper leading term and quasi-admissible projectors Q_0, …, Q_{k-1} such that G_k is nonsingular on $\mathcal{I} \subseteq \mathcal{J}$. We apply the decoupling procedure to this DAE as we used for regular DAEs in Chapter 2. Since $A = ADD^- = G_0D^-$, the DAE (2.17) can be rewritten as

$$G_0D^-(Dx)' + Bx \ = \ q. \tag{4.15}$$

The relations $x = P_0x + Q_0x = \Pi_0x + Q_0x$ gives

$$G_0D^-(Dx)' + B\Pi_0x + BQ_0x \ = \ q.$$

Using the properties $G_0D^- = G_1P_0D^- = G_1D^-$ and $BQ_0 = G_1Q_0$ we obtain

$$G_1D^-(Dx)' + B\Pi_0x + G_1Q_0x \ = \ q. \tag{4.16}$$

Similarly to the case of regular DAEs, this rearrangement can be extended to every level i in the chain. Namely, equation (4.15) can be transformed, via $G_0 = G_1P_0 = G_2P_1P_0 = \cdots = G_kP_{k-1}\cdots P_1P_0$, into

$$G_kP_{k-1}\cdots P_1P_0D^-(Dx)' + Bx \ = \ q. \tag{4.17}$$

Writing $B = B\Pi_{k-1} + B\Pi_{k-2}Q_{k-1} + \cdots + B\Pi_0Q_1 + BQ_0$, taking into account the definition of B_i, and applying the relations $B_iQ_i = G_kQ_i$, $G_i = G_kP_{k-1}\cdots P_i$, and

$$\begin{aligned} B\Pi_{i-1}Q_i \ &= \ B_iQ_i + \sum_{j=1}^{i} G_jD^-(D\Pi_jD^-)'D\Pi_{i-1}Q_i \\ &= \ G_kQ_i + \sum_{j=1}^{i} G_kP_{k-1}\cdots P_jD^-(D\Pi_jD^-)'D\Pi_{i-1}Q_i. \end{aligned}$$

for $i = 1, \ldots, k-1$, equation (4.17) is equivalent to

$$G_k P_{k-1} \cdots P_1 P_0 D^-(Dx)' + B\Pi_{k-1}x + \sum_{j=0}^{k-1} G_k Q_j x + \sum_{i=1}^{k-1} \sum_{j=1}^{i} G_k P_{k-1} \cdots P_j D^-(D\Pi_j D^-)' D\Pi_{i-1} Q_i x \;=\; q. \quad (4.18)$$

The derivatives $(D\Pi_j D^-)'$ start to appear for $k = 2$. Since G_k is nonsingular, we can scale (4.18) by G_k^{-1} to obtain

$$P_{k-1} \cdots P_1 P_0 D^-(Dx)' + G_k^{-1} B\Pi_{k-1}x + \sum_{j=0}^{k-1} Q_j x + \sum_{i=1}^{k-1} \sum_{j=1}^{i} P_{k-1} \cdots P_j D^-(D\Pi_j D^-)' D\Pi_{i-1} Q_i x \;=\; G_k^{-1} q. \quad (4.19)$$

As we have done for regular DAEs, we can decouple (4.19) into $k+1$ separate equations. Namely, multiplying (4.19) by $D\Pi_{k-1}$ and using the C^1 property of the projector $D\Pi_{k-1}D^-$ yield the inherent explicit regular ODE that determines the solution component $u := D\Pi_{k-1}x$ as depicted in (4.21). Further, if we multiply (4.19) by Q_{k-1} and then by Π_{k-2} if $k \geq 2$, we obtain the equation (4.22a). In turn, multiplying (4.19) by $Q_0 P_1 \cdots P_{k-1}$ and once again by $\Pi_{i-1} Q_i P_{i+1} \cdots P_{k-1}$, $i = 1, \ldots, k-2$, yields the algebraic relations defining the components $v_0 = Q_0 x$ and $v_i = \Pi_{i-1} Q_i x$, $i = 1, \ldots, k-1$, as shown in (4.22b) below. The following result is taken from [54].

Theorem 4.9. *Let equation* (2.17) *be a quasi-regular DAE on $\mathcal{I} \subseteq \mathcal{J}$. Then $x(t) \in C_D^1(\mathcal{I}, \mathbb{R}^m)$ is a solution of this DAE if it can be written as*

$$x \;=\; D^- u + v_0 + \cdots + v_{k-1}, \quad (4.20)$$

where $u = D\Pi_{k-1}x \in C^1(\mathcal{I}, \mathbb{R}^n)$ is a solution of the inherent explicit regular ODE

$$u' - (D\Pi_{k-1}D^-)'u + D\Pi_{k-1}G_k^{-1}BD^-u \;=\; D\Pi_{k-1}G_k^{-1}q, \quad (4.21)$$

whereas the components $v_0 = Q_0 x \in C(\mathcal{I}, \mathbb{R}^m)$ and $v_i = \Pi_{i-1}Q_i x \in C^1(\mathcal{I}, \mathbb{R}^m)$, $i = 1, \ldots, k-1$, satisfy

$$v_{k-1} = \mathcal{L}_{k-1}q - \mathcal{K}_{k-1}D^-u, \quad (4.22a)$$

$$v_i = \mathcal{L}_i q - \mathcal{K}_i D^- u + \sum_{j=i+1}^{k-1} \mathcal{N}_{ij}(Dv_j)' + \sum_{j=i+2}^{k-1} \mathcal{M}_{ij} v_j, \; i = 0, \ldots, k-2. \quad (4.22b)$$

The coefficients $\mathcal{L}_{k-1}$, $\mathcal{K}_{k-1}$ in (4.22a) *read*

$$\mathcal{L}_{k-1} \;=\; Q_{k-1}G_k^{-1}, \qquad \mathcal{K}_{k-1} = Q_{k-1}G_k^{-1}B\Pi_{k-1} \qquad \text{for } k = 1,$$

$$\mathcal{L}_{k-1} \;=\; \Pi_{k-2}Q_{k-1}G_k^{-1}, \quad \mathcal{K}_{k-1} = \Pi_{k-2}Q_{k-1}G_k^{-1}B\Pi_{k-1} \quad \textit{for } k \geq 2.$$

The coefficients $\mathcal{L}_i$, $\mathcal{K}_i$, $\mathcal{N}_{ij}$, *and* $\mathcal{M}_{ij}$ *in* (4.22b) *are given by*

$$\begin{aligned}
\mathcal{L}_0 &= Q_0P_1\cdots P_{k-1}G_k^{-1}, \quad \mathcal{L}_i \;=\; \Pi_{i-1}Q_iP_{i+1}\cdots P_{k-1}G_k^{-1}, \quad i = 1,\ldots,k-2,\\
\mathcal{N}_{01} &= Q_0Q_1D^-, \quad \mathcal{N}_{0j} \;=\; Q_0P_1\cdots P_{j-1}Q_jD^-, \quad j = 2,\ldots,k-1,\\
\mathcal{N}_{i,i+1} &= \Pi_{i-1}Q_iQ_{i+1}D^-, \quad i = 1,\ldots,k-2,\\
\mathcal{N}_{ij} &= \Pi_{i-1}Q_iP_{i+1}\cdots P_{j-1}Q_jD^-, \quad j = i+2,\ldots,k-1, \quad i = 1,\ldots,k-2,\\
\mathcal{K}_0 &= Q_0P_1\cdots P_{k-1}G_k^{-1}B\Pi_{k-1} + Q_0P_1\cdots P_{k-1}P_0D^-(D\Pi_{k-1}D^-)'D\Pi_{k-1},\\
\mathcal{K}_i &= \Pi_{i-1}Q_iP_{i+1}\cdots P_{k-1}G_k^{-1}B\Pi_{k-1}\\
&\quad + \Pi_{i-1}Q_iP_{i+1}\cdots P_{k-1}P_iD^-(D\Pi_{k-1}D^-)'D\Pi_{k-1}, \quad i = 1,\ldots,k-2,\\
\mathcal{M}_{ij} &= -\Pi_{i-1}Q_i\{Q_{i+1}D^-(D\Pi_iQ_{i+1}D^-)' + P_{i+1}Q_{i+2}D^-(D\Pi_{i+1}Q_{i+2}D^-)'\\
&\quad + \cdots + P_{i+1}\cdots P_{\mu-2}Q_{\mu-1}D^-(D\Pi_{\mu-2}Q_{\mu-1}D^-)'\}D\Pi_{j-1}Q_j\\
&\quad - \sum_{l=1}^{j}\Pi_{i-1}Q_iP_{i+1}\cdots P_{k-1}P_{k-1}\cdots P_lD^-(D\Pi_lD^-)'D\Pi_{j-1}Q_j,\\
&\quad j = i+2,\ldots,k-1, \quad i = 1,\ldots,k-2.
\end{aligned}$$

The proof of this result can be found in [54].

As we have seen, the quasi-admissible projector functions make it possible to decouple any quasi-regular DAE with quasi-proper leading term to obtain the system (4.21)–(4.22) which is similar to the one for regular DAEs. Nevertheless, the relationship between the values $r_i = \operatorname{rank} G_i$ and a possible index is impossible in case of quasi-admissible projectors. We refer to [54] for more detail .

Chapter 5

Index-2 DAEs with harmless critical points

Harmless critical points of linear time-varying DAEs

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \quad t \in \mathcal{J}, \tag{5.1}$$

with continuous matrix coefficients stated in Chapter 2 have been characterized under Assumptions 1 and 2 in Chapter 3, and according to the quasi-regular DAEs of Chapter 4. Under Assumptions 1 and 2, a critical point t_* of a linear index-2 DAE (5.1) is said to be harmless, if the resulting matrix $G_2(t)$ of the A-critical chain, cf. Section 3.2, is nonsingular for all $t \in \mathcal{J}$. It is clear that the type 2-A critical point can never be harmless since the matrix $G_2(t)$ has rank drop at t_*. Additionally, the critical point of type 1-A cannot be harmless because, as stated in Proposition 3.14, this type yields the singular matrix $G_2(t)$ at t_*. As a consequence, the harmless critical points, defined under Assumptions 1 and 2, of linear index-2 systems (5.1) can only be of type 0.

In case of quasi-regular setting, a critical point t_* of a linear quasi-regular DAE with $k = 2$ is called harmless, if the DAE having quasi-proper leading term is quasi-regular with nonsingular matrix $G_2(t)$ on $\mathcal{J}$ (cf. Definition 4.7). More precisely, there exist quasi-admissible projector functions Q_0, Q_1 such that $G_0(t)$ undergoes a rank deficiency at t_*, $G_1(t)$ has constant rank, and $G_2(t)$ is nonsingular for all $t \in \mathcal{J}$. As discussed in Chapter 4, the quasi-admissible projector functions generalize the projector functions constructed under Assumptions 1 and 2. We therefore define a linear index-2 DAE with harmless critical points as below.

Definition 5.1. *Equation* (5.1) *is called an index-2 DAE with harmless critical points on the interval $\mathcal{J}$ if the DAE is quasi-regular with the following conditions:*

(i) $G_0(t)$ *undergoes a rank deficiency at* t_**;*

(ii) $G_1(t)$ *has constant rank;*

(iii) $G_2(t)$ *is nonsingular for all* $t \in \mathcal{J}$*.*

In the previous chapters, the decoupling procedures have been discussed for linear regular and linear quasi-regular DAEs as well as linear DAEs with critical points.

As we are interested in the numerical solution of initial value problem of linear index-2 DAEs which holds harmless critical points, we address in this chapter the decoupling procedures of the linear index-2 problems (5.1) in more detail. The next section provides details of the decoupling method of the linear regular index-2 DAEs. This decoupling procedure is used to study stability and convergence properties of the integration methods for index-2 DAEs in Chapter 6. In Section 5.2 we discuss harmless critical points defined under Assumptions 1 and 2. The existence of harmless critical points makes it possible to define the canonical projector function Q_1 onto N_1 along S_1 as in case of regular setting. In particular, the scalarly implicit decoupling (3.8) and (3.9) formally coincides with the one in the regular problem, if the DAE possesses only harmless critical points. In Section 5.3 the decoupling of quasi-regular DAEs with $k = 2$ is given and shown to be equivalent to the one in the regular framework.

It should be mentioned that the calculation of these matrix functions and projector functions is only of theoretical interest. In numerical integration procedures, we do not need these special projector functions.

5.1 Decoupling of regular index-2 DAEs

Let (5.1) be a regular DAE with tractability index $\mu = 2$ on $\mathcal{I} \subseteq \mathcal{J}$ (cf. Definition 2.17 in Chapter 2). Thus, the matrix

$$G_2(t) = G_1(t) + B_1(t)Q_1(t)$$

is nonsingular for all $t \in \mathcal{I}$ with arbitrary admissible projector functions $Q_0(t)$, $Q_1(t)$ satisfying Definition 2.15. According to Lemma 2.4, we can choose $Q_1(t)$ to be the *canonical projector function* onto $\ker G_1(t)$ along $S_1(t) := \{z \in \mathbb{R}^m : B(t)z \in \operatorname{im} G_1(t)\}$ with the representation

$$Q_1(t) = Q_1(t)G_2^{-1}(t)B_1(t). \tag{5.2}$$

This Q_1 satisfies the relations

$$\begin{aligned} &Q_1Q_0 = Q_1G_2^{-1}B_1Q_0 = Q_1G_2^{-1}\left(BP_0 - G_1D^-(D\Pi_1D^-)'D\Pi_0\right)Q_0 = 0,\\ &Q_1G_2^{-1}B_1\Pi_1 = Q_1P_0P_1 = Q_1(I - Q_0)P_1 = 0, \end{aligned}$$

where the explicitly dependence in t has been removed. As stated in Section 2.3.2 of Chapter 2, by means of the projector-based method one can decompose the DAE (5.1) into 3 parts:

(1) the inherent explicit regular ODE (the dynamic part),

(2) part describing the inherent differentiation problem,

(3) algebraic part.

In the following, we drop the argument t for ease of notation. Applying the identities $A = ADD^- = G_2P_1P_0D^-$ and writing $B = B\Pi_1 + B\Pi_0Q_1 + BQ_0$, we can transform Equation (5.1) equivalently into the form

$$G_2P_1P_0D^-(Dx)' + B\Pi_1x + B\Pi_0Q_1x + BQ_0x = q. \tag{5.3}$$

The relations

$$B\Pi_0 = B_1 + G_1D^-(D\Pi_1D^-)'D\Pi_0,\ BQ_0 = G_2Q_0,\ B_1Q_1 = G_2Q_1,\ G_1 = G_2P_1,$$

and

$$\begin{aligned} B\Pi_0Q_1 &= B_1Q_1 + G_1D^-(D\Pi_1D^-)'D\Pi_0Q_1 \\ &= G_2Q_1 + G_2P_1D^-(D\Pi_1D^-)'D\Pi_0Q_1 \end{aligned} \tag{5.4}$$

imply

$$\begin{aligned} G_2P_1P_0D^-(Dx)' + B\Pi_1x + G_2P_1D^-(D\Pi_1D^-)'D\Pi_0Q_1x \\ + G_2Q_1x + G_2Q_0x = q. \end{aligned} \tag{5.5}$$

Scaling (5.5) by G_2^{-1} we get (cf. Equation (2.37))

$$P_1P_0D^-(Dx)' + G_2^{-1}B\Pi_1x + P_1D^-(D\Pi_1D^-)'D\Pi_0Q_1x + Q_1x + Q_0x = G_2^{-1}q. \tag{5.6}$$

Multiplying (5.6) by $D\Pi_1$ and using $\Pi_1Q_1 = 0$, $\Pi_1Q_0 = P_0\left(I - Q_1\right)Q_0 = 0$, lead to

$$D\Pi_1D^-(Dx)' + D\Pi_1G_2^{-1}B\Pi_1x + D\Pi_1D^-\left(D\Pi_1D^-\right)'D\Pi_0Q_1x = D\Pi_1G_2^{-1}q.$$

Due to $D\Pi_1D^-(D\Pi_1D^-)' = (D\Pi_1D^-)' - (D\Pi_1D^-)'D\Pi_1D^-$ and $\Pi_1\Pi_0Q_1 = 0$, this equation can be written to

$$D\Pi_1D^-(Dx)' + D\Pi_1G_2^{-1}B\Pi_1x + (D\Pi_1D^-)'D\Pi_0Q_1x = D\Pi_1G_2^{-1}q. \tag{5.7}$$

According to

$$D\Pi_1D^-(Dx)' = (D\Pi_1D^-Dx)' - (D\Pi_1D^-)'Dx = (D\Pi_1x)' - (D\Pi_1D^-)'D\Pi_0x,$$

Equation (5.7) is equivalent to

$$(D\Pi_1x)' - (D\Pi_1D^-)'D\Pi_0x + D\Pi_1G_2^{-1}B\Pi_1x + (D\Pi_1D^-)'D\Pi_0Q_1x = D\Pi_1G_2^{-1}q.$$

or

$$u' - (D\Pi_1D^-)'u + D\Pi_1G_2^{-1}BD^-u \ = \ D\Pi_1G_2^{-1}q, \tag{5.8}$$

where $D\Pi_0(I - Q_1)x = D\Pi_1x$ was used and $u := D\Pi_1x$. Equation (5.8) represents the so-called *inherent explicit regular ODE* for the component $u = D\Pi_1x$ of the linear regular index-2 DAE (5.1) and has the property that the solution belongs

to the invariant subspace $\operatorname{im} D\Pi_1$. That is, the solution starting in $\operatorname{im} D(t_0)\Pi_1(t_0)$ for some $t_0 \in \mathcal{I}$ remains in $\operatorname{im} D(t)\Pi_1(t)$ for all $t \in \mathcal{I}$.

In turn, multiplying (5.6) by $\Pi_0 Q_1$ and $Q_0 P_1$, respectively, we obtain, using $Q_1 P_1 = P_1 Q_1 = 0$, $Q_1 Q_0 = 0$, $Q_0 P_1 Q_0 = Q_0$,

$$\Pi_0 Q_1 G_2^{-1} B \Pi_1 x + \Pi_0 Q_1 x = \Pi_0 Q_1 G_2^{-1} q, \quad (5.9)$$

$$Q_0 P_1 D^- (Dx)' + Q_0 P_1 G_2^{-1} B \Pi_1 D^- D \Pi_1 x + Q_0 P_1 D^- (D\Pi_1 D^-)' D \Pi_0 Q_1 x + Q_0 x = Q_0 P_1 G_2^{-1} q. \quad (5.10)$$

Note that we have used $Q_0 P_1 G_2^{-1} B \Pi_1 x = Q_0 P_1 G_2^{-1} B \Pi_1 D^- D \Pi_1 x$ in (5.10). Taking into account

$$Q_1 = Q_1 G_2^{-1} B_1, \quad \Pi_1 \Pi_0 Q_1 = 0, \quad Q_0 P_1 \Pi_1 = 0, \quad Q_0 P_1 D^- = -Q_0 Q_1 D^-,$$

we find

$$\Pi_0 Q_1 G_2^{-1} B \Pi_1 x = \Pi_0 Q_1 G_2^{-1} \left(B_1 + G_2 P_1 D^- (D\Pi_1 D^-)' D \Pi_0 \right) P_1 x = 0, \quad (5.11)$$

$$\begin{aligned} Q_0 P_1 D^- (D\Pi_1 D^-)' D \Pi_0 Q_1 x &= Q_0 P_1 D^- ((\underbrace{D\Pi_1 D^- D Q_1 x}_{=0})' - D\Pi_1 D^- (DQ_1 x)') \\ &= -\underbrace{Q_0 P_1 D^- D \Pi_1}_{=0} D^- (D\Pi_0 Q_1 x)' = 0, \end{aligned} \quad (5.12)$$

as well as, using $(D\Pi_1 x)' = (D\Pi_1 D^- D \Pi_1 x)' = (D\Pi_1 D^-)' D\Pi_1 x + D\Pi_1 D^- (D\Pi_1 x)'$,

$$\begin{aligned} Q_0 P_1 D^- (Dx)' &= Q_0 P_1 D^- ((D\Pi_0 Q_1 x)' + (D\Pi_1 x)') \\ &= Q_0 P_1 D^- (D\Pi_0 Q_1 x)' + Q_0 P_1 D^- (D\Pi_1 D^-)' D \Pi_1 x \\ &\quad + Q_0 P_1 D^- D \Pi_1 D^- (D\Pi_1 x)' \\ &= -Q_0 Q_1 D^- (D\Pi_0 Q_1 x)' + Q_0 P_1 D^- (D\Pi_1 D^-)' D \Pi_1 D^- D \Pi_1 x. \end{aligned}$$

This makes it possible to formulate the system (5.9)-(5.10) to

$$\Pi_0 Q_1 x = \Pi_0 Q_1 G_2^{-1} q \quad (5.13)$$

$$Q_0 x = Q_0 P_1 G_2^{-1} q - Q_0 P_1 G_2^{-1} B \Pi_1 D^- D \Pi_1 x - Q_0 P_1 D^- (D\Pi_1 D^-)' D \Pi_1 D^- D \Pi_1 x + Q_0 Q_1 D^- (D\Pi_0 Q_1 x)'. \quad (5.14)$$

Denote $v_0 := Q_0 x$, $v_1 := \Pi_0 Q_1 x$, this system take the form

$$v_1 = \mathcal{L}_1 q, \quad (5.15)$$

$$v_0 = \mathcal{L}_0 q - \mathcal{K}_0 D^- u + \mathcal{N}_{01} (D v_1)', \quad (5.16)$$

where the continuous coefficients $\mathcal{L}_0$, $\mathcal{L}_1$, $\mathcal{K}_0$, $\mathcal{N}_{01}$ are given by

$$\mathcal{L}_1 = \Pi_0 Q_1 G_2^{-1}, \quad \mathcal{L}_0 = Q_0 P_1 G_2^{-1}, \quad \mathcal{N}_{01} = Q_0 Q_1 D^-$$
$$\mathcal{K}_0 = Q_0 P_1 G_2^{-1} B \Pi_1 + Q_0 P_1 D^- (D\Pi_1 D^-)' D \Pi_1.$$

Equation (5.15) is the *algebraic part* determining the component $v_1 = \Pi_0 Q_1 x$, whereas Equation (5.16) represents the *part describing the inherent differentiation problem* and makes it possible to find the component $v_0 = Q_0 x$; the so-called index-2 component. Thereby, we have to differentiate the component $v_1 = \Pi_0 Q_1 x$, or more precisely, we have to differentiate $\Pi_0 Q_1 G_2^{-1} q$.

Observe that the coefficient $\mathcal{K}_1 = \Pi_0 Q_1 G_2^{-1} B \Pi_1$ vanishes yielding the *complete* decoupling [66], due to the existence of the canonical projector $Q_1 = Q_1 G_2^{-1} B_1$ as depicted in (5.11). Further, the continuous coefficient $\mathcal{N}_{01} = Q_0 Q_1 D^-$ is nontrivial such that $\operatorname{im} Q_0 Q_1 = N_0 \cap S_0$ and the system (5.16) (resp. (5.14)) represents the differentiation problem.

Consequently, each solution x of the regular index-2 DAEs can be written as

$$\begin{aligned} x &= D^- u + v_0 + v_1 \\ &= (I - \mathcal{K}_0) D^- u + (\Pi_0 Q_1 + Q_0 P_1) G_2^{-1} q + Q_0 Q_1 D^- (D \Pi_0 Q_1 G_2^{-1} q)', \end{aligned}$$

where $(I - \mathcal{K}_0)$ is a nonsingular factor, $u = D\Pi_1 x \in C^1(\mathcal{I}, \mathbb{R}^n)$ is a solution of the inherent explicit regular ODE (5.8), and $v_1 = \Pi_0 Q_1 x$, $v_0 = Q_0 x$ are determined by (5.15), (5.16), respectively.

For the homogeneous index-2 DAE (5.1),

$$A(Dx)' + Bx = 0,$$

any solution x is given by

$$x = (I - \mathcal{K}_0) D^- u = (I - \mathcal{K}_0) D^- D \Pi_1 x = (I - \mathcal{K}_0) \Pi_1 D^- D \Pi_1 x = \Pi_{\text{can2}} D^- u,$$

where $\Pi_{\text{can2}} := (I - \mathcal{K}_0)\Pi_1$ is actually the projector function along the sum space $N_0 + N_1$. That is, by making use of the relation $\Pi_1 Q_0 = P_0 (I - Q_1) Q_0 = 0$, it results that

$$\begin{aligned} \Pi_{\text{can2}}^2 &= (I - \mathcal{K}_0)\Pi_1 (I - \mathcal{K}_0)\Pi_1 \\ &= (I - \mathcal{K}_0)\Pi_1 - (I - \mathcal{K}_0)\, \Pi_1 \left(Q_0 P_1 G_2^{-1} B \Pi_1 + Q_0 P_1 D^- (D \Pi_1 D^-)' D \Pi_1 \right) \Pi_1 \\ &= \Pi_{\text{can2}}, \end{aligned}$$

and, since $\ker \Pi_1 = N_0 + N_1$ and the factor $(I - \mathcal{K}_0)$ is nonsingular,

$$\ker \Pi_{\text{can2}} = \ker \Pi_1 = N_0 + N_1.$$

Also Π_{can2} is called the canonical projector function. One can approximate this projector as established in Theorem 5.2 below. The idea of this computation is originated from [82, 83] where it has been developed for the standard form DAE (1.3) with constant coefficients. Eventually, it will be useful for the derivation of error estimations for numerical integration methods in Chapter 6.

Theorem 5.2. *Let* (5.1) *be a linear regular DAE with tractability index 2 on* $\mathcal{I} \subseteq \mathcal{J}$. *If* $\eta > 0$ *is sufficiently small, then*

(1) *the matrix* $\left(\frac{1}{\eta}G_0 + B\right)$ *is nonsingular,*

(2) $$\left(\tfrac{1}{\eta}G_0 + B\right)^{-1} \tfrac{1}{\eta}G_0 = \Pi_{\text{can2}} + \Pi_{\text{can2}} \cdot O(\eta) \cdot \Pi_1 - \tfrac{1}{\eta}Q_0Q_1$$
$$-Q_0Q_1D^-(D\Pi_1D^-)'D\Pi_1 - Q_0Q_1D^-(D\Pi_1D^-)'D\Pi_1 \cdot O(\eta) \cdot \Pi_1, \quad (5.17)$$

(3) $\ker\left[\left(\frac{1}{\eta}G_0 + B\right)^{-1}\frac{1}{\eta}G_0\right]^2 = \ker \Pi_1$ *as* η *tends to zero*

Proof : (1) In order to prove the first statement we apply the above decoupling procedure to the equation

$$\left(\frac{1}{\eta}G_0 + B\right) z \;=\; w, \quad (5.18)$$

i.e., we use the identities $G_0 = G_2P_1P_0$ and $B = B\Pi_1 + B\Pi_0Q_1 + BQ_0$. The resulting equivalent equation of (5.18) reads

$$\frac{1}{\eta}G_2P_1P_0z + B\Pi_1z + B\Pi_0Q_1z + BQ_0z = w.$$

Then, we can reformulate this equation to

$$\frac{1}{\eta}G_2P_1P_0z + B\Pi_1z + G_2P_1D^-(D\Pi_1D^-)'D\Pi_0Q_1z + G_2Q_1z + BQ_0z = w, \quad (5.19)$$

by using the relations $BQ_0 = G_2Q_0$, $B\Pi_0Q_1 = G_2Q_1 + G_2P_1D^-(D\Pi_1D^-)'D\Pi_0Q_1$.

Scaling (5.19) by G_2^{-1} we get

$$\frac{1}{\eta}P_1P_0z + G_2^{-1}B\Pi_1z + P_1D^-(D\Pi_1D^-)'D\Pi_0Q_1z + Q_1z + Q_0z = G_2^{-1}w. \quad (5.20)$$

Multiplying (5.20) by $D\Pi_1$ and taking into account $\Pi_1z = \Pi_1D^-D\Pi_1z$ lead to

$$D\Pi_1z + \eta D\Pi_1G_2^{-1}B\Pi_1D^-D\Pi_1z + \eta D\Pi_1D^-(D\Pi_1D^-)'D\Pi_0Q_1z = \eta D\Pi_1G_2^{-1}w,$$

then to

$$D\Pi_1z = \left(I + \eta D\Pi_1G_2^{-1}B\Pi_1D^-\right)^{-1}\eta\left(D\Pi_1G_2^{-1}w\right.$$
$$\left. -\; D\Pi_1D^-(D\Pi_1D^-)'D\Pi_0Q_1z\right). \quad (5.21)$$

In turn, multiplying (5.20) by $\Pi_0 Q_1$ and $Q_0 P_1$, respectively, yields the system

$$\Pi_0 Q_1 G_2^{-1} B \Pi_1 z + \Pi_0 Q_1 z = \Pi_0 Q_1 G_2^{-1} w, \tag{5.22}$$

$$\frac{1}{\eta} Q_0 P_1 P_0 z + Q_0 P_1 G_2^{-1} B \Pi_1 z + Q_0 P_1 D^- (D\Pi_1 D^-)' D\Pi_0 Q_1 z + Q_0 z = Q_0 P_1 G_2^{-1} w. \tag{5.23}$$

The relation $Q_0 P_1 D^- (D\Pi_1 D^-)' D\Pi_0 Q_1 x = 0$ depicted in (5.12) and the existence of the canonical projector $Q_1 = Q_1 G_2^{-1} B_1$ allow us to write the system (5.22)-(5.23) as

$$\Pi_0 Q_1 z = \Pi_0 Q_1 G_2^{-1} w, \tag{5.24}$$

$$Q_0 z = Q_0 P_1 G_2^{-1} w - Q_0 P_1 G_2^{-1} B \Pi_1 D^- D \Pi_1 z + \frac{1}{\eta} Q_0 Q_1 \Pi_0 Q_1 z, \tag{5.25}$$

where $Q_0 P_1 P_0 = -Q_0 Q_1 \Pi_0 Q_1$ is applied. Consequently, we can write z as

$$\begin{aligned}
z &= D^- D\Pi_1 z + \Pi_0 Q_1 z + Q_0 z \\
&= D^- D\Pi_1 z + \Pi_0 Q_1 z + Q_0 P_1 G_2^{-1} w - Q_0 P_1 G_2^{-1} B \Pi_1 D^- D \Pi_1 z + \frac{1}{\eta} Q_0 Q_1 \Pi_0 Q_1 z \\
&= \left(I - Q_0 P_1 G_2^{-1} B \Pi_1\right) D^- D \Pi_1 z + \left(I + \frac{1}{\eta} Q_0 Q_1\right) \Pi_0 Q_1 z + Q_0 P_1 G_2^{-1} w \\
&= \left(I - Q_0 P_1 G_2^{-1} B \Pi_1\right) D^- \left(I + \eta D \Pi_1 G_2^{-1} B \Pi_1 D^-\right)^{-1} \eta \left(D \Pi_1 G_2^{-1} w \right. \\
&\quad \left. - D \Pi_1 D^- (D\Pi_1 D^-)' D \Pi_0 Q_1 z\right) + \left(I + \frac{1}{\eta} Q_0 Q_1\right) \Pi_0 Q_1 z \\
&\quad + Q_0 P_1 G_2^{-1} w \\
&= \left(I - Q_0 P_1 G_2^{-1} B \Pi_1\right) D^- \left(I + \eta D \Pi_1 G_2^{-1} B \Pi_1 D^-\right)^{-1} \eta \left(D \Pi_1 G_2^{-1} w \right. \\
&\quad \left. - D \Pi_1 D^- (D\Pi_1 D^-)' D \Pi_0 Q_1 G_2^{-1} w\right) + \left(I + \frac{1}{\eta} Q_0 Q_1\right) \Pi_0 Q_1 G_2^{-1} w \\
&\quad + Q_0 P_1 G_2^{-1} w.
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
\left(\frac{1}{\eta} G_0 + B\right)^{-1} = \left(I - Q_0 P_1 G_2^{-1} B \Pi_1\right) D^- \left(I + \eta D \Pi_1 G_2^{-1} B \Pi_1 D^-\right)^{-1} (D\Pi_1 \\
- D\Pi_1 D^- (D\Pi_1 D^-)' D\Pi_0 Q_1\Big) \eta G_2^{-1} \\
+ \left(\Pi_0 Q_1 + Q_0 P_1 + \frac{1}{\eta} Q_0 Q_1 \Pi_0 Q_1\right) G_2^{-1}
\end{aligned} \tag{5.26}$$

which proves the first statement.

(2) The properties $G_0 = G_2 P_1 P_0$, $Q_1 P_1 = 0$, and $Q_0 P_1 P_0 = -Q_0 Q_1$ imply

$$\begin{aligned}\left(\frac{1}{\eta}G_0 + B\right)^{-1}\frac{1}{\eta}G_0 &= \left(I - Q_0P_1G_2^{-1}B\Pi_1\right)D^-\left(I + \eta D\Pi_1G_2^{-1}B\Pi_1D^-\right)^{-1}D\Pi_1 \\ &\quad - \frac{1}{\eta}Q_0Q_1, \\ &= \left(I - Q_0P_1G_2^{-1}B\Pi_1\right)\left(D^-D\Pi_1 - \eta D^-D\Pi_1G_2^{-1}B\Pi_1D^-D\Pi_1\right. \\ &\quad \left. + \eta^2 D^-D\Pi_1G_2^{-1}B\Pi_1D^-D\Pi_1G_2^{-1}B\Pi_1D^-D\Pi_1 - \ldots\right) \\ &\quad - \frac{1}{\eta}Q_0Q_1 \\ &= \left(I - Q_0P_1G_2^{-1}B\Pi_1\right)\Pi_1\left(I - \eta G_2^{-1}B\Pi_1\right. \\ &\quad \left. + \eta^2 G_2^{-1}B\Pi_1G_2^{-1}B\Pi_1 - \ldots\right)\Pi_1 - \frac{1}{\eta}Q_0Q_1 \\ &= \left(I - Q_0P_1G_2^{-1}B\Pi_1\right)\Pi_1\left(I + \sum_{j=1}^{\infty}\left(-\eta G_2^{-1}B\Pi_1\right)^j\right) - \frac{1}{\eta}Q_0Q_1\end{aligned}$$

where we have made used of the relations

$$\begin{aligned}\left(I + \eta D\Pi_1G_2^{-1}B\Pi_1D^-\right)^{-1} &= \sum_{j=0}^{\infty}\left(-\eta D\Pi_1G_2^{-1}B\Pi_1D^-\right)^j \\ &= I - \eta D\Pi_1G_2^{-1}B\Pi_1D^- \\ &\qquad + \eta^2 D\Pi_1G_2^{-1}B\Pi_1D^-D\Pi_1G_2^{-1}B\Pi_1D^- - \ldots.\end{aligned}$$

Inserting $\Pi_{\text{can2}} = \left(I - Q_0P_1G_2^{-1}B\Pi_1 - Q_0P_1D^-(D\Pi_1D^-)'D\Pi_1\right)\Pi_1$ yields

$$\begin{aligned}\left(\frac{1}{\eta}G_0 + B\right)^{-1}\frac{1}{\eta}G_0 &= \Pi_{\text{can2}} + \Pi_{\text{can2}}\left(\sum_{j=1}^{\infty}\left(-\eta G_2^{-1}B\Pi_1\right)^j\right)\Pi_1 - \frac{1}{\eta}Q_0Q_1 \\ &\quad + Q_0P_1D^-(D\Pi_1D^-)'D\Pi_1 \\ &\quad + Q_0P_1D^-(D\Pi_1D^-)'D\Pi_1\left(\sum_{j=1}^{\infty}\left(-\eta G_2^{-1}B\Pi_1\right)^j\right)\Pi_1.\end{aligned}$$

Since $Q_0P_1D^- = -Q_0Q_1D^-$ and $\sum_{j=1}^{\infty}(-\eta G_2^{-1}B\Pi_1)^j = O(\eta)$ for $\eta \to 0$, the assertion is therefore verified.

(3) Due to $\Pi_1\Pi_{\text{can2}} = \Pi_1$, $\Pi_{\text{can2}}\Pi_1 = \Pi_{\text{can2}}$, and $\Pi_1Q_0 = 0$ it follows that

$$\begin{aligned}\left[\left(\frac{1}{\eta}G_0 + B\right)^{-1}\frac{1}{\eta}G_0\right]^2 &= \Pi_{\text{can2}} + \Pi_{\text{can2}}\cdot O(\eta)\cdot\Pi_1 - Q_0Q_1D^-(D\Pi_1D^-)'D\Pi_1 \\ &\quad - Q_0Q_1D^-(D\Pi_1D^-)'D\Pi_1\cdot O(\eta)\cdot\Pi_1. \qquad (5.27)\end{aligned}$$

($\Rightarrow$) From the relation (5.27) and $\left[\left(\frac{1}{\eta}G_0 + B\right)^{-1}\frac{1}{\eta}G_0\right]^2 z = 0$ we obtain

$$\begin{aligned} 0 \;=\; & \Pi_{\text{can2}} z + \Pi_{\text{can2}} \cdot O(\eta) \cdot \Pi_1 z - Q_0 Q_1 D^-(D\Pi_1 D^-)' D\Pi_1 z \\ & - Q_0 Q_1 D^-(D\Pi_1 D^-)' D\Pi_1 \cdot O(\eta) \cdot \Pi_1 z \end{aligned}$$

Multiplying this equation by Π_1 and applying $\Pi_1\Pi_{\text{can2}} = \Pi_1$, $\Pi_1 Q_0 = 0$ we find

$$\Pi_1 z + \Pi_1 \cdot O(\eta) \cdot \Pi_1 z = 0,$$

which immediately implies that $\Pi_1 z = 0$ if $\eta \to 0$.

($\Leftarrow$) Taking into account the relation (5.27) and $\Pi_1 z = 0$ we can prove that

$$\begin{aligned} \left[\left(\frac{1}{\eta}G_0 + B\right)^{-1}\frac{1}{\eta}G_0\right]^2 z &= \Pi_{\text{can2}} z \\ &= \left(I - Q_0 P_1 G_2^{-1} B\Pi_1 - Q_0 P_1 D^-(D\Pi_1 D^-)' D\Pi_1\right)\Pi_1 z = 0. \end{aligned}$$

The assertion (3) is therefore verified. □

Remark 5.3. *The matrix $(\frac{1}{\eta}G_0 + B)$ is evaluated and decomposed while processing the discretized equations. In case of the BDF method this matrix can be computed practically with $\eta := \frac{h_\ell}{\alpha_{0,l}}$. Therefore, the relation* (5.17) *turns out to be of practical use. Furthermore, the nonsingularity of the matrix $\left(\frac{1}{\eta}G_0 + B\right)$ verified in item (1) of Theorem 5.2 guarantees the feasibility of the BDF methods applied to the linear DAEs* (5.1)*. Let us consider again the index-2 problem in Example 1.3 For the implicit Euler method the matrix $\left(\frac{1}{\eta}G_0 + B\right)$ reads*

$$\left(\frac{1}{\eta}G_0 + B\right) = \left(\frac{1}{\eta}AD + B\right) = \frac{1}{h}\begin{bmatrix} 0 & 0 \\ 1 & \zeta t \end{bmatrix} + \begin{bmatrix} 1 & \zeta t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & \zeta t \\ \frac{1}{h} & \frac{\zeta}{h}t + 1 \end{bmatrix},$$

where $\eta = h$, $A = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $D = \begin{bmatrix} 1 & \zeta t \end{bmatrix}$, and $B = \begin{bmatrix} 1 & \zeta t \\ 0 & 1 \end{bmatrix}$. Obviously, due to $\det\left(\frac{1}{\eta}G_0 + B\right) = 1$, this matrix is nonsingular. On the other hand, if we apply the implicit Euler method to the standard formulation (1.9) *with $E = \begin{bmatrix} 0 & 0 \\ 1 & \zeta t \end{bmatrix}$ and $F = \begin{bmatrix} 1 & \zeta t \\ 0 & 1 + \zeta t \end{bmatrix}$ we obtain*

$$\left(\frac{1}{\eta}E + F\right) = \frac{1}{h}\begin{bmatrix} 0 & 0 \\ 1 & \zeta t \end{bmatrix} + \begin{bmatrix} 1 & \zeta t \\ 0 & 1 + \zeta t \end{bmatrix} = \begin{bmatrix} 1 & \zeta t \\ \frac{1}{h} & \frac{\zeta}{h}t + 1 + \zeta \end{bmatrix},$$

with $\det\left(\frac{1}{\eta}E + F\right) = \zeta + 1$, which is singular for $\zeta = -1$. This describes why the BDF method applied to the standard form DAE (1.9) *fails for $\zeta = -1$.*

5.2 Decoupling of index-2 DAEs with harmless critical points

Under the density hypothesis on the regular set stated in Assumption 1 and the requirements specified in Assumption 2 the linear DAEs with critical points can be described by means of the scalarly implicit decoupling (3.8) and (3.9) in Theorem 3.12, as indicated in Chapter 3. However, if the DAEs possess only harmless critical points, this decoupling procedure formally coincides with the one in the regular framework. In particular, we can choose $Q_1(t)$ as the canonical projector function onto $N_1(t)$ along $S_1(t)$ as in case of regular setting.

Proposition 5.4. *Let Assumptions 1 and 2 given in Section 3.2 hold. If the matrix G_2 is nonsingular for all $t \in \mathcal{J}$ and G_1 has constant rank, then $Q_1^c := Q_1 G_2^{-1} B_1$ is the canonical projector onto $N_1 = \ker G_1$ along $S_1 := \{z \in \mathbb{R}^m : B_1 z \in \operatorname{im} G_1\}$.*

Proof : We have to prove that $Q_1^c = Q_1 G_2^{-1} B_1$ is a projector function on $\mathcal{I}$ with

$$(Q_1^c)^2 = Q_1^c, \quad \operatorname{im} Q_1^c = N_1 \quad \text{and} \quad \ker Q_1^c = S_1.$$

(i) Q_1^c is a projector, because of

$$(Q_1^c)^2 = Q_1 G_2^{-1} B_1 Q_1 G_2^{-1} B_1 = Q_1 G_2^{-1} \underbrace{(G_1 + B_1 Q_1)}_{= G_2} Q_1 G_2^{-1} B_1 = Q_1^c.$$

(ii) The representation $Q_1^c = Q_1 G_2^{-1} B_1$ implies the relation $\operatorname{im} Q_1^c \subseteq \operatorname{im} Q_1$. Additionally, since the matrix function G_1 has constant rank on $\mathcal{I}$, we can choose Q_1 to be a projector onto $N_1 = \ker G_1$ with $\operatorname{im} Q_1 = N_1 = \ker G_1$ on $\mathcal{I}$. This yields

$$\operatorname{im} Q_1^c \ \subseteq \ \operatorname{im} Q_1 \ = \ N_1.$$

On the other hand, $z \in N_1 = \operatorname{im} Q_1$ means $z = Q_1 z = Q_1^2 z$. The nonsingularity of G_2 allows us to write z as, using the identity $G_1 Q_1 = 0$,

$$z = Q_1 (G_1 + B_1 Q_1)^{-1} (G_1 + B_1 Q_1) Q_1 z = Q_1 G_2^{-1} B_1 Q_1 z = Q_1^c Q_1 z = Q_1^c z,$$

that is, $z \in \operatorname{im} Q_1^c$.

(iii) $Q_1^c z = 0$ implies $(I - P_1) G_2^{-1} B_1 z = 0$. Then, we get $G_2^{-1} B_1 z = P_1 G_2^{-1} B_1 z$. Hence, due to $G_2 P_1 = G_1$, $B_1 z = G_2 P_1 G_2^{-1} B_1 z = G_1 G_2^{-1} B_1 z$, which shows that $B_1 z \in \operatorname{im} G_1$ meaning $z \in S_1$. Conversely, $z \in S_1$ indicates that there exists w such that $B_1 z = G_1 w$. Therefore, the identity

$$Q_1^c z = Q_1 G_2^{-1} B_1 z = Q_1 G_2^{-1} G_1 w = Q_1 G_2^{-1} G_2 P_1 w = 0$$

verifies $z \in \ker Q_1^c$. □

Now, let Assumptions 1 and 2 given in Chapter 3 hold and Equation (5.1) be a linear index-2 DAE with harmless critical points on $\mathcal{I} \subseteq \mathcal{J}$. Further, we choose

Q_1 as the canonical projector onto N_1 along S_1 with $Q_1 = Q_1 G_2^{-1} B_1$. Then, the scalarly implicit decoupling procedure (3.8) and (3.9) for the linear index-2 DAEs with harmless critical points is defined by

$$\omega_2 u' - \omega_2 (D\Pi_1 D^-)' u + D\Pi_1 G_2^{adj} B D^- u = D\Pi_1 G_2^{adj} q, \tag{5.28}$$

together with

$$\omega_2 v_1 = \mathcal{L}_1^{adj} q - \mathcal{K}_1^{adj} D^- u, \tag{5.29a}$$
$$\omega_2 v_0 = \mathcal{L}_0^{adj} q - \mathcal{K}_0^{adj} D^- u + \omega_2 \mathcal{N}_{01} (D v_1)', \tag{5.29b}$$

where $\omega_2 = \det G_2$ and G_2^{adj} is the transpose of the matrix of cofactors of G_2. The coefficients in (5.29) are given by

$$\begin{aligned}
\mathcal{L}_1^{adj} &= \Pi_0 Q_1 G_2^{adj}, \quad \mathcal{K}_1^{adj} = \Pi_0 Q_1 G_2^{adj} B \Pi_1, \\
\mathcal{L}_0^{adj} &= Q_0 P_1 G_2^{adj}, \quad \mathcal{N}_{01} = Q_0 Q_1 D^-, \\
\mathcal{K}_0^{adj} &= Q_0 P_1 G_2^{adj} B \Pi_1 + \omega_2 Q_0 P_1 D^- (D\Pi_1 D^-)' D \Pi_1.
\end{aligned}$$

As addressed in Definition 3.13, the harmlessness property leads to the nonsingularity of $G_2(t)$ for all $t \in \mathcal{I}$. Thus, the nonvanishing of $\omega_2(t) = \det G_2(t)$ makes it possible to divide the decoupling (5.28) and (5.29) by $\omega_2(t)$ yielding the system

$$u' - (D\Pi_1 D^-)' u + D\Pi_1 G_2^{-1} B D^- u = D\Pi_1 G_2^{-1} q, \tag{5.30a}$$
$$v_1 = \mathcal{L}_1 q - \mathcal{K}_1 D^- u, \tag{5.30b}$$
$$v_0 = \mathcal{L}_0 q - \mathcal{K}_0 D^- u + \mathcal{N}_{01} (D v_1)', \tag{5.30c}$$

with coefficients

$$\begin{aligned}
\mathcal{L}_1 &= \Pi_0 Q_1 G_2^{-1}, \quad \mathcal{K}_1 = \Pi_0 Q_1 G_2^{-1} B \Pi_1, \\
\mathcal{L}_0 &= Q_0 P_1 G_2^{-1}, \quad \mathcal{N}_{01} = Q_0 Q_1 D^-, \\
\mathcal{K}_0 &= Q_0 P_1 G_2^{-1} B \Pi_1 + Q_0 P_1 D^- (D\Pi_1 D^-)' D \Pi_1.
\end{aligned}$$

Furthermore, the existence of the canonical projector $Q_1 = Q_1 G_2^{-1} B_1$ implies $\mathcal{K}_1 = 0$, which simplifies (5.30b) to

$$v_1 = \mathcal{L}_1 q. \tag{5.31}$$

Therefore, each solution x can be written as

$$\begin{aligned}
x &= D^- u + v_0 + v_1 \\
&= (I - \mathcal{K}_0) D^- u + (\Pi_0 Q_1 + Q_0 P_1) G_2^{-1} q + Q_0 Q_1 D^- (D \Pi_0 Q_1 G_2^{-1} q)',
\end{aligned}$$

where $(I - \mathcal{K}_0)$ is a nonsingular factor, $u \in C^1(\mathcal{I}, \mathbb{R}^n)$ is a solution of the explicit ODE (5.30a) on the locally invariant space $\operatorname{im} D\Pi_1$, whereas v_0, v_1 satisfy (5.30c), (5.31), respectively.

Conversely, from the smoothness properties of u, v_0, and v_1 it follows that $x \in C_D^1(\mathcal{I}, \mathbb{R}^m)$, and additionally the relation $A(Dx)' + Bx = q$ is satisfied on $\mathcal{I} \cap \mathcal{J}_{reg}$. Since $A(Dx)' + Bx - q = 0$ is continuous and $\mathcal{I} \cap \mathcal{J}_{reg}$ is dense in $\mathcal{I}$, the identity $A(Dx)' + Bx = q$ has to hold in the whole $\mathcal{I}$. This proves that x actually solves the DAE (5.1) on $\mathcal{I}$.

Let us stress that the decoupling (5.30) for the linear index-2 DAEs with harmless critical points formally coincides with the one obtained for the regular index-2 problem in Section 5.1. Furthermore, as in case of regular setting, the relation $\ker \Pi_1 = N_0 + N_1$ is valid on the whole $\mathcal{I}$. This relies on the fact that, under Assumptions 1 and 2, the relation $\operatorname{im} Q_0(t) = N_0(t) \subseteq \ker G_0(t)$ holds for all t in the whole of $\mathcal{I}$ (cf. Remark 3.11). Since the matrix G_1 has constant rank on $\mathcal{I}$, we can choose Q_1 to be a projector onto $\ker G_1$ with $\operatorname{im} Q_1 = N_1 = \ker G_1$ on $\mathcal{I}$. Then, $P_0P_1z = 0$ means $z_0 := (I - Q_1)z \in N_0$, hence $z = Q_1z + z_0 \in N_0 + N_1$. On the other hand, due to $z \in N_0 + N_1$, we may decompose each z as $z = Q_0w_0 + Q_1w_1$. Then we compute $P_0P_1z = P_0P_1Q_0w_0 = P_0(I - Q_1)Q_0w_0 = 0$. Therefore $N_0 + N_1 \subseteq \ker P_0P_1$.

In addition, the canonical projector function Π_{can2} along the sum space $N_0 + N_1$ with the representation

$$\Pi_{\text{can2}} := (I - \mathcal{K}_0)\Pi_1 = \left(I - Q_0P_1G_2^{-1}B\Pi_1 - Q_0P_1D^-(D\Pi_1D^-)'D\Pi_1\right)\Pi_1$$

is well-defined on $\mathcal{I} \subseteq \mathcal{J}$. The result follows from the fact that $\Pi_{\text{can2}}^2 = \Pi_{\text{can2}}$ and Π_{can2} is continuous on $\mathcal{I}$, since all involved matrices defining it are continuous. Moreover, because of the nonsingularity of the factor $I - \mathcal{K}_0$ and $\ker \Pi_{\text{can2}} = \ker \Pi_1$, the relation

$$\ker \Pi_{\text{can2}} = \ker \Pi_1 = N_0 + N_1$$

is also satisfied in case of the index-2 DAEs with harmless critical points.

5.3 Decoupling of quasi-regular DAEs with k=2

For quasi-regular DAEs with $k = 2$ having the properties (i)-(iii) depicted above, the decoupling procedure (4.21) and (4.22) in Theorem 4.9 is given by

$$u' - (D\Pi_1D^-)'u + D\Pi_1G_2^{-1}BD^-u = D\Pi_1G_2^{-1}q, \tag{5.32a}$$

$$v_1 = \mathcal{L}_1q - \mathcal{K}_1D^-u, \tag{5.32b}$$

$$v_0 = \mathcal{L}_0q - \mathcal{K}_0D^-u + \mathcal{N}_{01}(Dv_1)', \tag{5.32c}$$

where $u = D\Pi_1x$, $v_1 = \Pi_0Q_1x$, and $v_0 = Q_0x$. The continuous coefficients in (5.32) read

$$\mathcal{L}_1 = \Pi_0Q_1G_2^{-1}, \quad \mathcal{K}_1 = \Pi_0Q_1G_2^{-1}B\Pi_1, \quad \mathcal{N}_{01} = Q_0Q_1D^-,$$

$$\mathcal{L}_0 = Q_0P_1G_2^{-1}, \quad \mathcal{K}_0 = Q_0P_1G_2^{-1}B\Pi_1 + Q_0P_1D^-(D\Pi_1D^-)'D\Pi_1.$$

Note that by quasi-regular DAEs with $k = 2$ one can additional define the canonical projector function Q_1 onto $N_1 = \ker G_1$ along $S_1 := \{z \in \mathbb{R}^m : B_1 z \in \operatorname{im} G_1\}$ with $Q_1 = Q_1 G_2^{-1} B_1$ to find

$$\mathcal{K}_1 = \Pi_0 Q_1 G_2^{-1} B \Pi_1 = \Pi_0 Q_1 G_2^{-1} \left(B_1 + G_2 P_1 D^- (D\Pi_1 D^-)' D\Pi_0 \right) P_1 x = 0.$$

Thus, Equation (5.32b) is simplified to

$$v_1 \;=\; \mathcal{L}_1 q. \tag{5.33}$$

Therefore, any solution x of the quasi-regular DAEs (5.1) with $k = 2$ can be expressed as

$$\begin{aligned} x &= D^- u + v_0 + v_1, \\ &= (I - \mathcal{K}_0) D^- u + (\Pi_0 Q_1 + Q_0 P_1) G_2^{-1} q + Q_0 Q_1 D^- (D\Pi_0 Q_1 G_2^{-1} q)', \end{aligned}$$

where $u \in C^1(\mathcal{I}, \mathbb{R}^n)$ is a solution of the inherent explicit regular ODE (5.32a), whereas $v_0 \in C(\mathcal{I}, \mathbb{R}^m)$, $v_1 \in C^1(\mathcal{I}, \mathbb{R}^m)$ satisfy (5.32c), (5.33), respectively.

As in case of regular DAEs, the canonical projector function Π_{can2} along the sum space $N_0 + N_1$ with

$$\Pi_{\text{can2}} := (I - \mathcal{K}_0)\Pi_1 = \left(I - Q_0 P_1 G_2^{-1} B \Pi_1 - Q_0 P_1 D^- (D\Pi_1 D^-)' D\Pi_1 \right) \Pi_1$$

is well-defined on $\mathcal{I} \subseteq \mathcal{J}$. The result follows from the fact that $\Pi_{\text{can2}}^2 = \Pi_{\text{can2}}$ and because of $\ker \Pi_{\text{can2}} = \ker \Pi_1$ and the nonsingularity of the factor $I - \mathcal{K}_0$, the relation

$$\ker \Pi_{\text{can2}} = \ker \Pi_1 = N_0 + N_1$$

is also satisfied in case of the quasi-regular DAEs.

As stated in [54], Equation (5.32a) is an explicit ODE for the component $u = D\Pi_1 x$, while Equation (5.33) represents an algebraic equation determining $v_1 = \Pi_0 Q_1 x$. Equation (5.32c) looks like a differentiation problem. However, in contrast to the regular DAEs, it may happen that the projector product $Q_0 Q_1$ vanishes on $\mathcal{I}$ or on subintervals, since now the subspaces N_0 do not necessarily coincide with $\ker G_0$. This is in the case if the DAE is quasi-regular with index 1 (see Definition 4.6). In this situation, we have the identity $Q_0 Q_1 = 0$ which yields the vanishing of the expression $\mathcal{N}_{01} = Q_0 Q_1 D^-$, and hence Equation (5.16) does not involve the differentiation problem anymore.

Let us consider again the quasi-regular DAE (4.3) in Example 4.1 (writing in standard form as (4.1)). The matrix function sequence defined by the quasi-admissible projector functions is computed in Example 4.8. There, the projector product $Q_0 Q_1$ reads

$$Q_0 Q_1 = \begin{bmatrix} 0 & -\alpha \\ 0 & 0 \end{bmatrix},$$

which implies that $Q_0Q_1 = 0$, if $\alpha(t) = 0$. As addressed in Example 4.1, on the interval where $\alpha(t) = 0$, the DAE (4.1) is regular with index $\mu = 1$, whereas on the interval where $\alpha(t) \neq 0$, the DAE (4.1) is regular with index $\mu = 2$.

Chapter 6

Numerical integrations of index-2 DAE with harmless critical points

Numerical integration methods for explicit regular ODEs have been studied in numerous papers. *One step methods* such as *Runge-Kutta* methods were first proposed by Runge [80] and Heun [39], and was completely characterized by Kutta [51] for the set of fourth order methods. The application of the Runge-Kutta methods for the numerical solution of DAEs is developed in [33, 35, 49, 52, 78]. *Backward differentiation formulas* (BDF) belonging to *linear multistep methods* have been first stated by Curtiss and Hirschfelder [16], but their particular application to stiff equations has been recognized since the work of Gear [22]. As a consequence, several codes based on BDF methods for differential-algebraic systems were written [6, 11, 79]. The well-known code DASSL of Petzold [72], was implemented for index-1 DAEs and was described in detail in [10]. Further implementations are LSODI of Hindmarsh [44] and SPRINT of Berzins and Furzeland [5]. The BDF schemes for numerical solving DAEs have also been considered in [7, 8, 10, 27, 33, 49, 55].

For DAEs with properly stated leading term the numerical integrations were studied in [41, 42, 43, 63]. Stability and convergence results of the implicit Runge-Kutta (IRK) and BDF approaches applied to nonlinear index-1 properly stated DAE

$$A(x(t),t)(Dx)'(t) + b(x(t),t) = q(t), \qquad t \in \mathcal{J},$$

were first reported by Higueras and März in [41]. It was proved that if the subspace $\operatorname{im} D(t)$ is constant, then the numerical method integrates the inherent regular ODE associated with the problem. In [43] the results on qualitative properties of the numerical solution of linear index-2 DAEs of the form

$$A(t)(Dx)'(t) + B(t)x(t) = q(t), \qquad t \in \mathcal{J}, \tag{6.1}$$

with properly stated leading term are given under the assumption that the two subspaces associated with the index-2 DAE, say DN_1 and DS_1, are constant. In the same cite it was shown that the IRK and BDF methods applied to (6.1) are *weakly instable* meaning that in the stability inequality there exists an additional term, which is proportional to $\frac{1}{h}$ [60], on the right-hand side.

In this chapter we adapt the standard integration methods originally developed for

regular ODEs to solve an index-2 DAE (6.1) possessing harmless critical points on a compact interval $\mathcal{I} = [t_0, T] \subseteq \mathcal{J}$. In particular, we study the stability and convergence properties of IRK and BDF methods applied to our problems. Here, we follow the line of [41, 43]. Recall from Definition 5.1 that a linear index-2 DAE (6.1) with harmless critical points has the following properties:

(i) G_0 has a rank drop at critical points,

(ii) G_1 has constant rank on $\mathcal{J}$,

(iii) G_2 is nonsingular for all $t \in \mathcal{J}$,

where the matrix functions G_i define under Assumptions 1–2 in Chapter 3 and according to the quasi-regular DAEs in Chapter 4.

Let $x_* \in C_D^1(\mathcal{I}, \mathbb{R}^m)$ be a solution of the DAE (6.1) to be approximated and let the interval $\mathcal{I} := [t_0, T] \subseteq \mathcal{J}$ be partitioned by

$$\pi : t_0 < t_1 < \cdots < t_\ell < \cdots < t_N = T \tag{6.2}$$

with the stepsize $h_\ell := t_\ell - t_{\ell-1}$, $\ell = 1, \ldots, N$. Further, let h_{max} be the maximal stepsize of the grid π, $h_{max} := \max_{\ell=1,\ldots,N} h_\ell$.

6.1 Runge-Kutta Methods

In this section, we deal with a Runge-Kutta method applied to the index-2 DAE holding harmless critical points. Runge-Kutta methods are one step methods. These are methods which use only one initial approximation $x_{\ell-1} \approx x_*(t_{\ell-1})$ at the beginning of a step in order to compute the approximation x_ℓ at the current time point t_ℓ. We will start with the application of the Runge-Kutta methods to explicit regular ODEs and then will show how these schemes apply to the systems of DAEs.

Consider an initial value problem for ordinary differential systems of the form

$$x'(t) = F(x(t), t), \qquad x(t_0) = x^0, \tag{6.3}$$

on the compact interval $\mathcal{I} = [t_0, T]$. An s-stage Runge-Kutta method applied to (6.3) reads

$$x_\ell = x_{\ell-1} + h \sum_{i=1}^{s} \beta_i X'_{\ell i}, \tag{6.4}$$

where the stage derivatives $X'_{\ell i}$ are computed from

$$X'_{\ell i} = F(X_{\ell i}, t_{\ell i}), \qquad i = 1, \ldots, s, \tag{6.5}$$

and the stage approximations $X_{\ell i}$ of the solutions $x_*(t_{\ell i})$ at the stages $t_{\ell i}$ are given

by

$$X_{\ell i} \;=\; x_{\ell-1} + h\sum_{j=1}^{s} \alpha_{ij} X'_{\ell j}, \qquad i = 1, \ldots, s. \tag{6.6}$$

Here $x_{\ell-1}$ is an approximation to the solution $x_*(t_{\ell-1})$ and $t_{\ell i} = t_{\ell-1} + c_i h, i = 1, \ldots, s$, are intermediate stages, $c_i \in [0, 1]$. The coefficients α_{ij}, β_i, c_i determine the Runge-Kutta method, and s is the number of stages. The Runge-Kutta method is called *implicit*, if there exist α_{ij} with $\alpha_{ij} \neq 0$ and $i \geq j$. In general, the method's parameters can conveniently be represented in a so-called Butcher tableau

$$\begin{array}{c|cccc} c_1 & \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1s} \\ c_2 & \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & \alpha_{s1} & \alpha_{s2} & \cdots & \alpha_{ss} \\ \hline & \beta_1 & \beta_2 & \cdots & \beta_s \end{array} \quad = \quad \begin{array}{c|c} c & \mathcal{A} \\ \hline & \beta^T \end{array}$$

with coefficients $\mathcal{A} := (\alpha_{ij})_{i,j=1}^{s}$, $\beta := (\beta_1, \beta_2, \ldots, \beta_s)^T$, $c := (c_1, c_2, \ldots, c_s)^T$, and $c_i = \sum_{j=1}^{s} \alpha_{ij}, i = 1, \ldots, s$. Observe that (6.4) and (6.6) depend on the method and only (6.5) depends on the equation (6.3).

In order to construct a Runge-Kutta method that has a certain consistency of order p, meaning that the error per step can be estimated in terms of $O(h^{p+1})$, the coefficients α_{ij}, β_i, c_i of the method need to satisfy the simplifying assumptions [13]

$$\begin{aligned} B(p): &\quad \sum_{i=1}^{s} \beta_i c_i^{k-1} &&= \tfrac{1}{k}, && k = 1, \ldots, p, \\ C(\eta): &\quad \sum_{j=1}^{s} \alpha_{ij} c_j^{k-1} &&= \tfrac{1}{k} c_i^k, && k = 1, \ldots, \eta, \quad i = 1, \ldots, s, \\ D(\zeta): &\quad \sum_{i=1}^{s} \beta_i c_i^{k-1} \alpha_{ij} &&= \tfrac{1}{k} \beta_j (1 - c_j^k), && k = 1, \ldots, \zeta, \quad j = 1, \ldots, s. \end{aligned} \tag{6.7}$$

Condition $B(p)$ is necessary for a method of order p. If the requirement $C(q)$ holds, then the method has stage order q, i.e., the stage approximations are computed with accuracy $|X^*_{\ell i} - x_*(t_{\ell-1} + c_i h)| = O(h^{q+1})$, where $X^*_{\ell i}$ is the approximate solution obtained from the exact initial value $x_\ell = x_*(t_\ell)$. The importance of these conditions is summarized in the following fundamental result [12, 13].

Theorem 6.1. *If the coefficients α_{ij}, β_i, c_i of a Runge-Kutta method applied to (6.3) satisfy $B(p)$, $C(\eta)$, $D(\zeta)$ with $p \leq \eta + \zeta + 1$ and $p \leq 2\eta + 2$, then the method is of order p, that is, $|x^*_\ell - x_*(t_\ell)| = O(h^{p+1})$ if x^*_ℓ is the numerical solution calculated by (6.4) using the exact initial value $x_\ell = x_*(t_\ell)$ and stepsize h.*

One may construct a Runge-Kutta method using the assumptions (6.7). For instance, the Gauss methods, which is based on the Gaussian quadrature formulas,

1	1	$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$	$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$	$\frac{296-169\sqrt{6}}{1800}$	$\frac{-2+3\sqrt{6}}{225}$
	1	1	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\sqrt{6}}{1800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$
			$\frac{3}{4}$	$\frac{1}{4}$	1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
						$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
$s=1$		$s=2$			$s=3$			

Table 6.1: Some examples of Radua IIA methods

are derived by first choosing $c_1, \ldots, c_s$ to be the zeros of the shifted Legendre polynomials of degree s,

$$p_s(t) := \frac{d^s}{dt^s}(t^s(t-1)^s),$$

and then choosing the coefficient matrix $\mathcal{A}$ and β such that $B(2s)$ and $C(s)$ are satisfied. These methods have stage order s and order $2s$, which is the highest possible order for an s-stage Runge-Kutta method. Another important family of Runge-Kutta methods is based on the Radau quadrature schemes. The parameters $c_1, \ldots, c_s$ are chosen to be the zeros of the polynomial $p_s - p_{s-1}$ with $c_1 = 0$ or $c_s = 1$. Choosing $c_s = 1$ and β and $\mathcal{A}$ such that $B(s)$ and $C(s)$ are satisfied, yields the Radau IIA schemes [13, 19, 33]. The order of a Radau IIA method is $p = 2s-1$ and the stage order is $q = s$. Some examples of Radua IIA methods are given in Table 6.1.

For the nonlinear DAE of the form

$$f(x'(t), x(t), t) \;=\; 0, \tag{6.8}$$

with a consistent initial value $x(t_0) = x_0$, the Runge-Kutta methods can be realized in the following way [10]. The numerical approximation x_ℓ of the solution $x_*(t)$ at time point t_ℓ is obtained from

$$x_\ell = x_{\ell-1} + h\sum_{i=1}^{s} \beta_i X'_{\ell i}, \tag{6.9}$$

where the stage derivatives $X'_{\ell i}$ are defined by

$$f\left(X'_{\ell i}, X_{\ell i}, t_{\ell i}\right) \;=\; 0, \qquad i = 1, \ldots, s, \tag{6.10}$$

and the stage approximations $X_{\ell i}$ are determined from

$$X_{\ell i} = x_{\ell-1} + h\sum_{j=1}^{s} \alpha_{ij} X'_{\ell j}, \qquad i = 1, \ldots, s. \tag{6.11}$$

Note that Equation (6.5) in the Runge-Kutta scheme is replaced by Equation (6.10) if the ODE (6.3) is replaced by the DAE (6.8). Since the matrix $\frac{\partial f}{\partial x'}$ is

singular, certain components of the stage derivatives have to be computed from (6.11). Obviously, a condition for $X'_{\ell 1}, \ldots, X'_{\ell s}$ to be uniquely defined by (6.10), (6.11), is the nonsingularity of the coefficient matrix $\mathcal{A}$ [56, 81]. Therefore, we shall assume the coefficient matrix $\mathcal{A}$ to be nonsingular in the sequel.

Let $(\tilde{\alpha}_{ij})_{i,j=1}^{s}$ denote the elements of the matrix $\mathcal{A}^{-1}$, that is, $\mathcal{A}^{-1} := (\tilde{\alpha}_{ij})_{i,j=1}^{s}$, and $\otimes$ be the Kronecker product of the matrices [18, 45]. Once the stage approximations $X_{\ell i}$ are computed, we can solve (6.11) for $X'_{\ell i}$ from

$$\sum_{j=1}^{s} \alpha_{ij} X'_{\ell j} \;=\; \frac{1}{h}\left(X_{\ell i} - x_{\ell-1}\right), \qquad i = 1, \ldots, s.$$

The coefficient matrix of this system is $\mathcal{A} \otimes I_m$, being nonsingular, and $(\mathcal{A} \otimes I_m)^{-1} = \mathcal{A}^{-1} \otimes I_m$. The system solution is then given by

$$X'_{\ell i} \;=\; \frac{1}{h} \sum_{j=1}^{s} \tilde{\alpha}_{ij} (X_{\ell j} - x_{\ell-1}), \qquad i = 1, \ldots, s. \tag{6.12}$$

Therefore, with the notations $X_\ell = (X_{\ell 1}, \ldots, X_{\ell s})^T$ and $\mathbb{1} := (1, \ldots, 1)^T \in \mathbb{R}^s$, we can write the approximation x_ℓ as

$$\begin{aligned} x_\ell \;&=\; x_{\ell-1} + h \sum_{i=1}^{s} \beta_i \frac{1}{h} \sum_{j=1}^{s} \tilde{\alpha}_{ij} (X_{\ell j} - x_{\ell-1}) \\ &=\; \left(1 - \sum_{i=1}^{s}\sum_{j=1}^{s} \beta_i \tilde{\alpha}_{ij}\right) x_{\ell-1} + \sum_{i=1}^{s}\sum_{j=1}^{s} \beta_i \tilde{\alpha}_{ij} X_{\ell j} \\ &=\; \underbrace{\left(1 - \beta^T \mathcal{A}^{-1} \mathbb{1}\right)}_{=:\,\rho} x_{\ell-1} + \left(\beta^T \mathcal{A}^{-1} \otimes I_m\right) X_\ell. \end{aligned}$$

It is well known that the IRK methods with

$$\mathcal{A} \text{ nonsingular}, \quad c_s = 1, \quad \text{and} \quad \beta_i = \alpha_{si}, \quad i = 1, \ldots, s, \tag{6.13}$$

are an appropriate tool to handle DAEs [27, 73]. This special class of IRK schemes yields $\rho = 0$, $x_\ell = X_{\ell s}$, and $t_{\ell s} = t_{\ell-1} + h = t_\ell$ guaranteeing that the approximation x_ℓ generated by the Runge-Kutta methods satisfies the constraint. That is, for each solution $x_*(\cdot)$ of the DAE (6.8) the relation $x_*(t) \in \mathcal{M}_0(t)$ with

$$\mathcal{M}_0(t) \;:=\; \{x \in \mathcal{D}_f : \exists y \in \mathbb{R}^n : f(y, x, t) = 0\}, \tag{6.14}$$

holds. Similarly, according to equation (6.10) the stage approximations $X_{\ell i}$ belong to the set $\mathcal{M}_0(t)$, i.e., $X_{\ell i} \in \mathcal{M}_0(t_{\ell i})$ for $i = 1, \ldots, s$. As stated above, assuming $c_s = 1$ and $\beta_i = \alpha_{si}$ for $i = 1, \ldots, s$, we have $x_\ell = X_{\ell s}$, and $t_{\ell s} = t_{\ell-1} + h = t_\ell$. Thus,

$$x_\ell \;=\; X_{\ell s} \;\in\; \mathcal{M}_0(t_{\ell s}) \;=\; \mathcal{M}_0(t_\ell)$$

imply that the approximation x_ℓ satisfies the constrain of the system.

Therefore, an IRK method satisfying property (6.13) simplifies (6.9), (6.11) to

$$\begin{aligned} x_\ell &= X_{\ell s}, \\ X'_{\ell i} &= \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(X_{\ell j} - x_{\ell-1}), \qquad i = 1, \ldots, s. \end{aligned}$$

The Radau IIA methods are examples of the special IRK methods [33]. This special class of the IRK methods has been applied to the quasi-regular DAEs in [40, 41, 43].

Definition 6.2. *An implicit Runge-Kutta method with nonsingular matrix $\mathcal{A}$ is said to be* stiffly accurate *if the last row of $\mathcal{A}$ coincides with the vector β^T, that is, the condition $\beta_i = \alpha_{si}$, $i = 1, \ldots, s$, is satisfied.*

Consequently, we will use the IRK methods satisfying (6.13) to approximate the solution of the linear DAE (6.1). Given an approximation $x_{\ell-1}$ to the solution $x_*(t_{\ell-1})$ and a stepsize h, we solve the system

$$A_{\ell i}[DX]'_{\ell i} + B_{\ell i}X_{\ell i} = q_{\ell i}, \qquad i = 1, \ldots, s, \tag{6.15}$$

for $X_{\ell 1}, \ldots, X_{\ell s}$ where the stage derivatives $[DX]'_{\ell i}$ are defined by

$$[DX]'_{\ell i} = \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(D_{\ell j}X_{\ell j} - D_{\ell-1}x_{\ell-1}), \qquad i = 1, \ldots, s. \tag{6.16}$$

Then, an approximation x_ℓ of the exact solution $x_*(t_\ell) = x_*(t_{\ell-1} + h)$ is given by

$$x_\ell = X_{\ell s}. \tag{6.17}$$

Here, $A_{\ell i} := A(t_{\ell i})$, $D_{\ell i} := D(t_{\ell i})$, $B_{\ell i} := B(t_{\ell i})$, $q_{\ell i} := q(t_{\ell i})$, $i = 1, \ldots, s$.

6.1.1 Convergence result

In order to investigate the stability and convergence properties of the Runge-Kutta methods (6.15) the decoupling procedure, introduced in Chapter 5, is applied to the discretized equation (6.15). That is, we use the relations

$$\begin{aligned} A_{\ell i} &= A_{\ell i}D_{\ell i}D^-_{\ell i} = G_{2,\ell i}P_{1,\ell i}P_{0,\ell i}D^-_{\ell i} \\ B_{\ell i} &= B_{\ell i}\Pi_{1,\ell i} + B_{\ell i}\Pi_{0,\ell i}Q_{1,\ell i} + B_{\ell i}Q_{0,\ell i} \\ &= B_{\ell i}\Pi_{1,\ell i} + G_{2,\ell i}P_{1,\ell i}D^-_{\ell i}\hat{R}'_{\ell i}D_{\ell i}\Pi_{0,\ell i}Q_{1,\ell i} + G_{2,\ell i}Q_{1,\ell i} + G_{2,\ell i}Q_{0,\ell i}, \end{aligned}$$

and transform the difference equation (6.15) into

$$\begin{aligned} G_{2,\ell i}P_{1,\ell i}P_{0,\ell i}D^-_{\ell i}[DX]'_{\ell i} + B_{\ell i}\Pi_{1,\ell i}X_{\ell i} + G_{2,\ell i}P_{1,\ell i}D^-_{\ell i}\hat{R}'_{\ell i}D_{\ell i}\Pi_{0,\ell i}Q_{1,\ell i}X_{\ell i} \\ + G_{2,\ell i}Q_{1,\ell i}X_{\ell i} + G_{2,\ell i}Q_{0,\ell i}X_{\ell i} = G_{2,\ell i}q_{\ell i}, \end{aligned} \tag{6.18}$$

where the notation $\hat{R} := D\Pi_1 D^-$ is introduced to simplify expressions. Then, we scale equation (6.18) by $G_{2,\ell i}^{-1}$ and multiply the resulting system by $D_{\ell i}\Pi_{1,\ell i}$, $\Pi_{0,\ell i}Q_{1,\ell i}$, and $Q_{0,\ell i}P_{1,\ell i}$, respectively. Further reformulations (cf. (5.7),(5.10),(5.9)) yield the decoupled system of (6.15):

$$\hat{R}_{\ell i}[DX]'_{\ell i} + (D\Pi_1 G_2^{-1} B\Pi_1)_{\ell i} X_{\ell i} + \hat{R}'_{\ell i} D_{\ell i}\Pi_{0,li}Q_{1,li}X_{\ell i} = (D\Pi_1 G_2^{-1} q)_{\ell i} \quad (6.19a)$$

$$-(Q_0 Q_1 D^-)_{\ell i}[DX]'_{\ell i} + (Q_0\Pi_1 G_2^{-1} B\Pi_1)_{\ell i} X_{\ell i} + Q_{0,li}X_{\ell i} = (Q_0\Pi_1 G_2^{-1} q)_{\ell i} \quad (6.19b)$$

$$\Pi_{0,\ell i}Q_{1,li}X_{\ell i} = (\Pi_0 Q_1 G_2^{-1} q)_{\ell i}. \quad (6.19c)$$

Introducing $[MX]'_{\ell i} = \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(M(t_{\ell j})X_{\ell j} - M(t_{\ell-1})x_{\ell-1})$ for $i = 1, \ldots, s$ and any matrix function $M(t)$ defined on $\mathcal{J}$ to derive

$$\begin{aligned}
[DX]'_{\ell i} &= \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(D_{\ell j}X_{\ell j} - D_{\ell-1}x_{\ell-1}) \\
&= \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(D_{\ell j}\Pi_{0,\ell j}P_{1,\ell j}X_{\ell j} + D_{\ell j}\Pi_{0,\ell j}Q_{1,\ell j}X_{\ell j} \\
&\qquad - D_{\ell-1}\Pi_{0,\ell-1}P_{1,\ell-1}x_{\ell-1} - D_{\ell-1}\Pi_{0,\ell-1}Q_{1,\ell-1}x_{\ell-1}), \\
&= [D\Pi_1 X]'_{\ell i} + [D\Pi_0 Q_1 X]'_{\ell i},
\end{aligned}$$

where the relations $D = D\Pi_0$ and $I = P_1 + Q_1$ were used.

Consequently, denoting $U_{\ell i} := D_{\ell i}\Pi_{1,\ell i}X_{\ell i}$, and $V_{\ell i} := \Pi_{0,\ell i}Q_{1,\ell i}X_{\ell i}$, the decoupled system (6.19) can be rewritten as

$$\hat{R}_{\ell i}U'_{\ell i} + (D\Pi_1 G_2^{-1} BD^-)_{\ell i}U_{\ell i} + \hat{R}_{\ell i}[DV]'_{\ell i} + [\hat{R}]'_{\ell i}D_{\ell i}V_{\ell i} = (D\Pi_1 G_2^{-1} q)_{\ell i} \quad (6.20a)$$

$$-(Q_0 Q_1 D^-)_{\ell i}[DV]'_{\ell i} - (Q_0 Q_1 D^-)_{\ell i}U'_{\ell i} + (Q_0 P_1 G_2^{-1} BD^-)_{\ell i}U_{\ell i} + Q_{0,li}X_{\ell i} = (Q_0 P_1 G_2^{-1} q)_{\ell i} \quad (6.20b)$$

$$V_{\ell i} = (\Pi_0 Q_1 G_2^{-1} q)_{\ell i}. \quad (6.20c)$$

Following the approach proposed in [41], we consider a perturbed system with perturbations $\delta_{\ell i}^{[1]}$ and $\delta_{\ell i}^{[2]}$:

$$A_{\ell i}[DX^{[1]}]'_{\ell i} + B_{\ell i}X_{\ell i}^{[1]} - q_{\ell i} \;=\; \delta_{\ell i}^{[1]}, \qquad i = 1, \ldots, s \quad (6.21)$$

where $[DX^{[1]}]'_{\ell i} := \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(D_{\ell j}X_{\ell j}^{[1]} - D_{\ell-1}x_{\ell-1}^{[1]})$, $i = 1, \ldots, s$, and

$$A_{\ell i}[DX^{[2]}]'_{\ell i} + B_{\ell i}X_{\ell i}^{[2]} - q_{\ell i} \;=\; \delta_{\ell i}^{[2]}, \qquad i = 1, \ldots, s \quad (6.22)$$

where $[DX^{[2]}]'_{\ell i} := \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(D_{\ell j}X_{\ell j}^{[2]} - D_{\ell-1}x_{\ell-1}^{[2]})$, $i = 1, \ldots, s$.

Let $\Delta_{\ell i} := \delta_{\ell i}^{[1]} - \delta_{\ell i}^{[2]}$. Subtracting the decoupled system of (6.22) from the decoupled system of (6.21) leads to

$$\hat{R}_{\ell i}\left(U_{\ell i}^{[1]}-U_{\ell i}^{[2]}\right)'+(D\Pi_1 G_2^{-1}BD^-)_{\ell i}\left(U_{\ell i}^{[1]}-U_{\ell i}^{[2]}\right)$$
$$+\hat{R}_{\ell i}\left([DV^{[1]}]'_{\ell i}-[DV^{[2]}]'_{\ell i}\right)+\hat{R}'_{\ell i}D_{\ell i}\left(V_{\ell i}^{[1]}-V_{\ell i}^{[2]}\right)=(D\Pi_1 G_2^{-1}\Delta)_{\ell i} \quad (6.23a)$$
$$-(Q_0Q_1D^-)_{\ell i}\left([DV^{[1]}]'_{\ell i}-[DV^{[2]}]'_{\ell i}\right)-(Q_0Q_1D^-)_{\ell i}\left(U_{\ell i}^{[1]}-U_{\ell i}^{[2]}\right)'$$
$$+(Q_0P_1G_2^{-1}BD^-)_{\ell i}\left(U_{\ell i}^{[1]}-U_{\ell i}^{[2]}\right)+Q_{0,li}(X_{\ell i}^{[1]}-X_{\ell i}^{[2]})=(Q_0P_1G_2^{-1}\Delta)_{\ell i} \quad (6.23b)$$
$$V_{\ell i}^{[1]}-V_{\ell i}^{[2]}=(\Pi_0Q_1G_2^{-1}\Delta)_{\ell i}. \quad (6.23c)$$

Below we show that the IRK method, when applied to linear index-2 DAEs with harmless critical points, becomes unstable. However, this instability is weak in the sense that certain components are amplified by $\frac{1}{h}$ and this error is not accumulated.

Theorem 6.3. *Let $x_* \in C_D^1([t_0,T],\mathbb{R}^m)$ be a solution of the index-2 DAE* (6.1) *with harmless critical points (cf. Definition 5.1). Let the IRK method with the property* (6.13) *be given and let the maximal stepsizes of all partitions* (6.2) *be sufficiently small. Then, for all sufficiently small perturbations $|\delta_{\ell i}^{[1]}|\le\delta$, $|\delta_{\ell i}^{[2]}|\le\delta$, $\delta>0$, and $x_\ell^{[1]}$ and $x_\ell^{[2]}$ associated via* (6.21), (6.22), *respectively, the following stability estimate holds,*

$$\|x_\ell^{[1]}-x_\ell^{[2]}\| \le K\left(\|D_0x_0^{[1]}-D_0x_0^{[2]}\|+\max_{k\le\ell}\max_{1\le i\le s}\|\delta_{ki}^{[1]}-\delta_{ki}^{[2]}\|\right.$$
$$\left.+\max_{k\le\ell}\max_{1\le i\le s}\left\{\frac{1}{h}\|(D\Pi_0Q_1G_2^{-1})_{ki}(\delta_{ki}^{[1]}-\delta_{ki}^{[2]})\|\right\}\right), \quad \ell\ge 1, \quad (6.24)$$

with a constant $K>0$ independent of the stepsize.

Proof : Denoting

$$E_{\ell i}:=U_{\ell i}^{[1]}-U_{\ell i}^{[2]},\quad i=1,\dots,s,\quad e_{\ell-1}:=u_{\ell-1}^{[1]}-u_{\ell-1}^{[2]},$$
$$F_{\ell i}:=V_{\ell i}^{[1]}-V_{\ell i}^{[2]},\quad i=1,\dots,s,\quad f_{\ell-1}:=v_{\ell-1}^{[1]}-v_{\ell-1}^{[2]},$$

system (6.23) can be written as

$$\hat{R}_{\ell i}E'_{\ell i}+(D\Pi_1G_2^{-1}BD^-)_{\ell i}E_{\ell i}+\hat{R}_{\ell i}[DF]'_{\ell i}+\hat{R}'_{\ell i}D_{\ell i}F_{\ell i}=(D\Pi_1G_2^{-1}\Delta)_{\ell i} \quad (6.25a)$$
$$-(Q_0Q_1D^-)_{\ell i}[DF]'_{\ell i}-(Q_0Q_1D^-)_{\ell i}E'_{\ell i}$$
$$+(Q_0P_1G_2^{-1}BD^-)_{\ell i}E_{\ell i}+Q_{0,li}(X_{\ell i}^{[1]}-X_{\ell i}^{[2]})=(Q_0P_1G_2^{-1}\Delta)_{\ell i} \quad (6.25b)$$
$$F_{\ell i}=(\Pi_0Q_1G_2^{-1}\Delta)_{\ell i}, \quad (6.25c)$$

where

$$E'_{\ell i}:=\frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(E_{\ell j}-e_{\ell-1}),\qquad i=1,\dots,s,$$

$$[DF]'_{\ell i}:=\frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(D_{\ell j}F_{\ell j}-D_{\ell-1}f_{\ell-1}),\qquad i=1,\dots,s.$$

Due to $\hat{R}_{\ell i}E_{\ell i} = E_{\ell i}$ and $\hat{R}_{\ell i}D_{\ell i}F_{\ell i} = 0$, for $i = 1, \ldots, s$, it follows that

$$\hat{R}_{\ell i}E'_{\ell i} = E'_{\ell i} + \frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(\hat{R}_{\ell i} - \hat{R}_{\ell j})E_{\ell j} - \frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(\hat{R}_{\ell i} - \hat{R}_{\ell-1})e_{\ell-1}, \tag{6.26}$$

$$\hat{R}_{\ell i}[DF]'_{\ell i} = \frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(\hat{R}_{\ell i} - \hat{R}_{\ell j})D_{\ell j}F_{\ell j} - \frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(\hat{R}_{\ell i} - \hat{R}_{\ell-1})D_{\ell-1}f_{\ell-1}. \tag{6.27}$$

Then, inserting (6.25c), (6.26), and (6.27) into (6.25a) and denoting

$$\begin{aligned}
\tilde{W}_{\ell i} &:= -\Pi_{1,\ell i}G_{2,\ell i}^{-1}B_{\ell i}D_{\ell i}^{-},\\
\varphi_{\ell i} &:= \left(D_{\ell i}\Pi_{1,\ell i}G_{2,\ell i}^{-1} - \hat{R}'_{\ell i}D_{\ell i}\Pi_{0,\ell i}Q_{1,\ell i}G_{2,\ell i}^{-1}\right)\Delta_{\ell i}\\
\phi_{\ell i} &:= \frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(\hat{R}_{\ell i} - \hat{R}_{\ell j})E_{\ell j} - \frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(\hat{R}_{\ell i} - \hat{R}_{\ell-1})e_{\ell-1},\\
\psi_{\ell i} &:= \frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(\hat{R}_{\ell i} - \hat{R}_{\ell j})D_{\ell j}F_{\ell j} - \frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}(\hat{R}_{\ell i} - \hat{R}_{\ell-1})D_{\ell-1}f_{\ell-1},
\end{aligned}$$

we can write (6.25a) as

$$\begin{bmatrix} E'_{\ell 1}\\ \vdots\\ E'_{\ell s}\end{bmatrix} = \begin{bmatrix} D_{\ell 1}\tilde{W}_{\ell 1} & & \\ & \ddots & \\ & & D_{\ell s}\tilde{W}_{\ell s}\end{bmatrix}\begin{bmatrix} E_{\ell 1}\\ \vdots\\ E_{\ell s}\end{bmatrix} + \begin{bmatrix} \varphi_{\ell 1}\\ \vdots\\ \varphi_{\ell s}\end{bmatrix} - \begin{bmatrix} \phi_{\ell 1}\\ \vdots\\ \phi_{\ell s}\end{bmatrix} - \begin{bmatrix} \psi_{\ell 1}\\ \vdots\\ \psi_{\ell s}\end{bmatrix}. \tag{6.28}$$

With the abbreviations $\mathcal{D}_M = \operatorname{diag}(M(t_{\ell 1}), \ldots, M(t_{\ell s}))$ for any matrix function $M(t)$ defined on $\mathcal{J}$, $E_\ell = (E_{\ell 1}, \ldots, E_{\ell s})^T$, $E'_\ell = (E'_{\ell 1}, \ldots, E'_{\ell s})^T$, and in a similar way for φ_ℓ, ϕ_ℓ, and ψ_ℓ, Equation (6.28) can be written in the compact form

$$E'_\ell = \mathcal{D}_{D\tilde{W}}E_\ell + \varphi_\ell - \phi_\ell - \psi_\ell.$$

Applying the Runge-Kutta recursion

$$hE'_\ell = (\mathcal{A}^{-1} \otimes I)(E_\ell - \mathbb{1} \otimes e_{\ell-1})$$

yields

$$((\mathcal{A}^{-1} \otimes I) - h\mathcal{D}_{D\tilde{W}})E_\ell = (\mathcal{A}^{-1} \otimes I)(\mathbb{1} \otimes e_{\ell-1}) + h\varphi_\ell - h\phi_\ell - h\psi_\ell,$$

and hence,

$$(\mathcal{A} \otimes I)^{-1}(\mathcal{A} \otimes I)((\mathcal{A}^{-1} \otimes I) - h\mathcal{D}_{D\tilde{W}})E_\ell = (\mathcal{A}^{-1} \otimes I)(\mathbb{1} \otimes e_{\ell-1}) + h\varphi_\ell - h\phi_\ell - h\psi_\ell.$$

Since $(\mathcal{A}\otimes I)(\mathcal{A}^{-1}\otimes I)=I$ it follows that

$$(\mathcal{A}\otimes I)^{-1}(I-h(\mathcal{A}\otimes I)\mathcal{D}_{D\tilde{W}})E_\ell=(\mathcal{A}^{-1}\otimes I)(\mathbb{1}\otimes e_{\ell-1})+h\varphi_\ell-h\phi_\ell-h\psi_\ell.$$

Due to the regularity of the coefficient matrix $\mathcal{A}$ and the continuity of $D\Pi_1G_2^{-1}BD^-$, there exists an h_* such that the matrix $(I-h(\mathcal{A}\otimes I)\mathcal{D}_{D\tilde{W}})$ is nonsingular for $h\le h_*$ and

$$\|(I-h(\mathcal{A}\otimes I)\mathcal{D}_{D\tilde{W}})^{-1}\| \;\le\; 1+hC_1,$$

for some constant $C_1>0$ independent of the stepsize. Consequently, we receive the estimate

$$\|E_\ell\|\le(1+hC_1)\|e_{\ell-1}\|+hC_2\|\varphi_\ell\|+hC_2\|\phi_\ell\|+hC_2\|\psi_\ell\|. \tag{6.29}$$

According to condition (iv) of Assumption 2, the projector function $\hat{R}$ is continuously differentiable on the compact interval $\mathcal{I}=[t_0,T]$. Then, we have

$$\|\phi_\ell\|\le L(\|E_\ell\|+\|e_{\ell-1}\|)\quad\text{and}\quad\|\psi_\ell\|\le L(\|D_\ell F_\ell\|+\|D_{\ell-1}f_{\ell-1}\|).$$

Hence, (6.29) reads

$$\begin{aligned}(1-hC_2L)\|E_\ell\| \;\le\; &(1+hC_1+hC_2L)\|e_{\ell-1}\|+hC_2\|\varphi_\ell\|\\ &+hC_2L(\|D_\ell F_\ell\|+\|D_{\ell-1}f_{\ell-1}\|).\end{aligned} \tag{6.30}$$

Thus, for $h\le\frac{1}{2C_2L}$, (6.30) gives

$$\|E_\ell\| \;\le\; (1+hC_3)\|e_{\ell-1}\|+hC_4\|\varphi_\ell\|+hC_5(\|D_\ell F_\ell\|+\|D_{\ell-1}f_{\ell-1}\|). \tag{6.31}$$

Due to the continuity of $D\Pi_1G_2^{-1}$ and the continuous differentiability of $\hat{R}$, it follows that

$$\|\varphi_\ell\|\le K_1\max_{1\le i\le s}\|\Delta_{\ell i}\|.$$

Furthermore, because of the continuity of $D\Pi_0Q_1G_2^{-1}$, (6.25c) provides

$$\|D_\ell F_\ell\|\le K_2\max_{1\le i\le s}\|\Delta_{\ell i}\|\quad\text{and}\quad\|D_{\ell-1}f_{\ell-1}\|\le K_3\|\Delta_{\ell-1,s}\|.$$

Therefore, the estimate (6.31) can be rewritten as

$$\|E_\ell\| \;\le\; (1+hC_3)\|e_{\ell-1}\|+hC_6\max_{\ell-1\le k\le\ell}\max_{1\le i\le s}\|\Delta_{ki}\|. \tag{6.32}$$

Using $\|e_\ell\|=\|E_{\ell s}\|\le\|E_\ell\|$ the estimate for $\|e_\ell\|$ reads

$$\|e_\ell\| \;\le\; (1+hC_3)\|e_{\ell-1}\|+hC_6\max_{\ell-1\le k\le\ell}\max_{1\le i\le s}\|\Delta_{ki}\|.$$

Now, the standard recursion process provides the required stability bound for e_ℓ,

$$\|e_\ell\| \;\le\; (1+hC_3)^\ell\|e_0\| + hC_6\sum_{k=0}^{\ell-1}(1+hC_3)^k \max_{k\le\ell}\max_{1\le i\le s}\|\Delta_{ki}\|.$$

Since

$$h\sum_{k=0}^{\ell-1}(1+hC_3)^k \;=\; \frac{(1+hC_3)^\ell-1}{C_3} \;\le\; \frac{(1+hC_3)^\ell}{C_3},$$

and $(1+hC_3)^\ell \le \exp^{C_3\ell h} = \exp^{C_3(T-t_0)}$, we obtain

$$\|e_\ell\| \;\le\; \exp^{C_3(T-t_0)}\left(\|e_0\| + \frac{C_6}{C_3}\max_{k\le\ell}\max_{1\le i\le s}\|\Delta_{ki}\|\right).$$

Consequently, there exists a constant $K_e > 0$ such that

$$\|e_\ell\| \;\le\; K_e\left(\|e_0\| + \max_{k\le\ell}\max_{1\le i\le s}\|\Delta_{ki}\|\right).$$

From $x_\ell = X_{\ell s} = D_{\ell s}^- U_{\ell s} + V_{\ell s} + Q_{0,\ell s}X_{\ell s}$ it follows that

$$\begin{aligned}
x_\ell^{[1]} - x_\ell^{[2]} &= D_{\ell s}^- E_{\ell s} + F_{\ell s} + Q_{0,\ell s}\left(X_{\ell s}^{[1]} - X_{\ell s}^{[2]}\right)\\
&= (D_{\ell s}^- - (Q_0P_1G_2^{-1}BD^-)_{\ell s})E_{\ell s} + \left((\Pi_0Q_1G_2^{-1})_{\ell s} + (Q_0P_1G_2^{-1})_{\ell s}\right)\Delta_{\ell s}\\
&\quad + (Q_0Q_1D^-)_{\ell s}\frac{1}{h}\sum_{i=1}^{s}\tilde\alpha_{si}(D\Pi_0Q_1G_2^{-1})_{\ell i}\,(\Delta_{\ell i} - \Delta_{\ell-1})\\
&\quad + (Q_0Q_1D^-)_{\ell s}((-D\Pi_1G_2^{-1}BD^-)_{\ell s}E_{\ell s} + \varphi_{\ell s} - \phi_{\ell s} - \psi_{\ell s}). \qquad (6.33)
\end{aligned}$$

Finally, we can estimate

$$\begin{aligned}
\|x_\ell^{[1]} - x_\ell^{[2]}\| \le& K_4\|e_\ell\| + K_5\|\Delta_{\ell s}\|\\
&+ K_6\max_{1\le i\le s}\left(\frac{1}{h}\|(D\Pi_0Q_1G_2^{-1})_{\ell i}\|\,(\|\Delta_{\ell i}\| + \|\Delta_{\ell-1}\|)\right)\\
&+ K_6\,(\|\varphi_{\ell s}\| + \|\phi_{\ell s}\| + \|\psi_{\ell s}\|)\\
\le& K\left(\|e_0\| + \max_{k\le\ell}\max_{1\le i\le s}\|\Delta_{ki}\| + \max_{k\le\ell}\max_{1\le i\le s}\left(\frac{1}{h}\|(D\Pi_0Q_1G_2^{-1})_{ki}(\Delta_{ki})\|\right)\right)
\end{aligned}$$

and receive the required estimation. □

It is important to note that the expression $Q_0Q_1D^-$ in equation (6.33) vanishes, if the DAE is quasi-regular with index 1 (see Definition 4.6). In this case the error estimation (6.24) simplifies to

$$\|x_\ell^{[1]} - x_\ell^{[2]}\| \;\le\; K\left(\|D_0x_0^{[1]} - D_0x_0^{[2]}\| + \max_{k\le\ell}\max_{1\le i\le s}\|\delta_{ki}^{[1]} - \delta_{ki}^{[2]}\|\right), \quad \ell \ge 1. \quad (6.34)$$

Now we can state the convergence result for IRK method when applied to linear index-2 DAE (6.1) with harmless critical points.

Theorem 6.4. *Let $x_* \in C_D^1([t_0,T],\mathbb{R}^m)$ be a solution of the index-2 DAE* (6.1) *with harmless critical points (cf. Definition 5.1), and let the IRK method with the property* (6.13) *be given. If the IRK method satisfies the order condition $C(q)$ and $x_*(\cdot)$ satisfies $D(\cdot)x_*(\cdot) \in C^{q+1}$, then the IRK method* (6.15) *is convergent with order q.*

Proof : Consider the approximation x_ℓ solved by the IRK method

$$A_{\ell i}[DX]'_{\ell i} + B_{\ell i}X_{\ell i} - q_{\ell i} = 0, \qquad i = 1,\dots,s.$$

Following [34], we define the local error $\tau_{\ell i}$ of the IRK method as

$$\tau_{\ell i} \;:=\; A_{\ell i}[Dx_*]'_{\ell i} + B_{\ell i}x_*(t_{\ell i}) - q_{\ell i}, \qquad i = 1,\dots,s,$$

where

$$[Dx_*]'_{\ell i} \;:=\; \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(D_{\ell j}x_*(t_{\ell j}) - D_{\ell-1}x_*(t_{\ell-1})), \qquad i = 1,\dots,s.$$

The local error $\tau_{\ell i}$ is of order $O(h^q)$, because of the order condition $C(q)$. Further, $\tau_{\ell i}$ lies in $\operatorname{im} A_{\ell i}$. This follows from the fact that the exact solution $x_*(t_{\ell i})$ of (6.1) satisfies

$$A_{\ell i}(Dx_*)'(t_{\ell i}) + B_{\ell i}x_*(t_{\ell i}) - q_{\ell i} = 0, \qquad i = 1,\dots,s,$$

which make it possible to write

$$\begin{aligned}\tau_{\ell i} \;&=\; A_{\ell i}[Dx_*]'_{\ell i} - A_{\ell i}(Dx_*)'(t_{\ell i}), \qquad i = 1,\dots,s,\\ &=\; A_{\ell i}\left(\frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}\left(D_{\ell j}x_*(t_{\ell j}) - D_{\ell-1}x_*(t_{\ell-1})\right) - (Dx_*)'(t_{\ell i})\right), \qquad i = 1,\dots,s.\end{aligned}$$

Due to Theorem 6.3, we obtain

$$\begin{aligned}\|x_\ell - x_*(t_\ell)\| \;\le\; K\Big(&\|D_0x_0 - D_0x_*(t_0)\| \;+\; \max_{k\le\ell}\max_{1\le i\le s}\|\tau_{ki}\| \\ &+ \max_{k\le\ell}\max_{1\le i\le s}\left(\frac{1}{h}\|(D\Pi_0Q_1G_2^{-1})_{ki}\tau_{ki}\|\right)\Big).\end{aligned}$$

Since $\tau_{\ell i} \in \operatorname{im} A_{\ell i}$, there exists a w such that $\tau_{\ell i} = A_{\ell i}w, i = 1,\dots,s$. This implies $G_{2,\ell i}^{-1}\tau_{\ell i} = P_{1,\ell i}P_{0,\ell i}D_{\ell i}^{-}w$ which leads to

$$D_{\ell i}\Pi_{0,\ell i}Q_{1,\ell i}G_{2,\ell i}^{-1}\tau_{\ell i} \;=\; D_{\ell i}\Pi_{0,\ell i}Q_{1,\ell i}P_{1,\ell i}P_{0,\ell i}D_{\ell i}^{-}w \;=\; 0, \qquad i = 1,\dots,s.$$

Then, we have

$$\|x_\ell - x_*(t_\ell)\| \le K\left(\|D_0x_0 - D_0x_*(t_0)\| + \max_{k\le\ell}\max_{1\le i\le s}\|\tau_{ki}\|\right).$$

Therefore, the method is convergent with order q. □

6.2 Backward Differentiation Formula

BDF method is a class of implicit linear multistep methods that is particularly useful for stiff differential equations and differential-algebraic systems. In contrast to the Runge-Kutta schemes presented in Section 6.1 the BDF methods require starting values from several previous integration steps and need fewer function evaluations per step in order to compute the approximation x_ℓ of the solution $x_*(t)$ at time point t_ℓ.

We consider again the initial value problem for ODE of the form (6.3) and assume the approximations $x_{\ell-j}$ to the solutions $x_*(t_{\ell-j})$, $j = 0, 1, \ldots, k$, to be given. A BDF method can be derived by differentiating the polynomial $p(t)$ of degree k which interpolates the values $\{(t_{\ell-j}, x_{\ell-j}) : j = 0, 1, \ldots, k\}$ and setting the derivative $p'(t)$ at time point t_ℓ to $f(x_\ell, t_\ell)$, that is, $p'(t_\ell) = f(x_\ell, t_\ell)$. This gives the formula of the k-step BDF method

$$\frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} x_{\ell-j} \;=\; f(x_\ell, t_\ell), \qquad \ell \geq k, \tag{6.35}$$

where $\alpha_{\ell,j}$, $j = 0, 1, \ldots, k$, $\alpha_{\ell,0} \neq 0$, are the coefficients of the method and $h_\ell := t_\ell - t_{\ell-1}$ is the current stepsize. For $k = 1$ this is the implicit Euler method. The k-step constant stepsize BDF method for ODEs is stable for $k \leq 6$ and unstable for $k > 6$ [13, 15, 29, 36]. An introduction to the properties of BDF methods for ODEs can be found in [23, 37, 53].

Lemma 6.5. *Let $y(\cdot) \in C^{k+2}[t_0, T]$, $p(\cdot)$ denote the polynomial of degree k which interpolates the values $\{(t_{\ell-j}, y(t_{\ell-j})) : j = 0, 1, \ldots, k\}$, and $\frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} y(t_{\ell-j}) = p'(t_\ell)$ be k-step backward difference to y at time point t_ℓ. Then*

$$y'(t_\ell) - \frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} y(t_{\ell-j}) \;=\; \frac{1}{(k+1)!} c_{\ell,k} y^{(k+1)}(t_\ell) h_\ell^k + O(h_\ell^{k+1}), \tag{6.36}$$

where $c_{\ell,k}$ is the computable constant.

Proof : Since $p(t)$ interpolates the given function $y(t)$ at the nodes $t_\ell, \ldots, t_{\ell-k}$, we get the interpolation error

$$y(t) - p(t) \;=\; \prod_{j=0}^{k} (t - t_{\ell-j})\, y[t_{\ell-k}, \ldots, t_\ell, t],$$

where $y[t_{\ell-k}, \ldots, t_\ell, t]$ is the notation for the divided differences. Then,

$$y'(t) \;=\; p'(t) + \sum_{j=0}^{k} \prod_{i \neq j} (t - t_{\ell-i})\, y[t_{\ell-k}, \ldots, t_\ell, t] + \prod_{j=0}^{k} (t - t_{\ell-j})\, (y[t_{\ell-k}, \ldots, t_\ell, t])'$$

$$
\begin{aligned}
y'(t_\ell) &= p'(t_\ell) + \sum_{j=0}^{k} \prod_{i \neq j} (t_\ell - t_{\ell-i})\, y[t_{\ell-k}, \ldots, t_\ell, t_\ell] \\
&= p'(t_\ell) + \underbrace{\prod_{i \neq 0} (t_\ell - t_{\ell-i})}_{= c_{\ell,k} h_\ell^k} \underbrace{y[t_{\ell-k}, \ldots, t_\ell, t_\ell]}_{= \frac{1}{(k+1)!} y^{(k+1)}(\zeta_\ell),\, \zeta_\ell \in [t_{\ell-k}, t_\ell]} .
\end{aligned}
$$

With $y^{(k+1)}(\zeta_\ell) = y^{(k+1)}(t_\ell) + O(h_\ell)$ the statement is proved. □

For the nonlinear DAE of the form (6.8) with a consistent initial value $x(t_0) = x_0$ the k-step BDF methods, $k \leq 6$, can be constructed by approximating the derivative $x'(t)$ in (6.8) by a backward difference quotient $\frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} x_{\ell-j}$. This yields the system of nonlinear equations

$$
f\left(\frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} x_{\ell-j}, x_\ell, t_\ell \right) = 0, \qquad \ell \geq k, \tag{6.37}
$$

which is usually solved for x_ℓ at each time step by Newton's method. Obviously, the numerical solution x_ℓ generated by the BDF methods belongs to the constraint set $\mathcal{M}_0(t_\ell)$ of the system.

A k-step BDF method ($k \leq 6$) applied to the linear time-invariant DAE (6.1) reads

$$
A_\ell \frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_{\ell-j} + B_\ell x_\ell = q_\ell, \qquad \ell \geq k, \tag{6.38}
$$

where $A_\ell := A(t_\ell)$, $D_\ell := D(t_\ell)$, $B_\ell := B(t_\ell)$, and $q_\ell := q(t_\ell)$. Let

$$
[Dx]'_\ell := \frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_{\ell-j}.
$$

Then, equation (6.38) can be written simply as

$$
A_\ell [Dx]'_\ell + B_\ell x_\ell = q_\ell, \qquad \ell \geq k. \tag{6.39}
$$

The next subsection deals with the study of the stability and convergence properties of the BDF method applied to the linear index-2 DAE (6.1) with harmless critical points.

6.2.1 Convergence result

A convergence analysis of BDF schemes for linear constant coefficient DAEs was first given in [84]. Convergence results for the BDF method applied to nonlinear index-1 DAE systems (6.8) was presented e.g. in [8, 10, 26, 27, 55]. For semi-explicit index-2 problems

$$
y' = f(y, z), \qquad 0 = g(y),
$$

the stability and convergence properties of the BDF are also described in [33]. The stability property of the BDF schemes for quasi-linear index-2 DAEs (arising from circuit simulation) of the form

$$A(t)x'(t) + g(x,t) \;=\; 0, \tag{6.40}$$

has been studied in [90, 91] under the assumption that the nullspace $\ker A$ and the space $N(x,t) \cap S(x,t)$ are constant. The analysis of the stability and convergence of BDF methods applied to quasi-linear DAEs

$$A(x,t)(d(x,t)') + b(x,t) \;=\; 0$$

was presented in [41, 43]. We extend this concept in order to study convergence and stability properties of the BDF schemes when applied to the index-2 DAE (6.1) with harmless critical points. We consider the partitions π defined in (6.2) with the following properties:

$$h_{min} \le h_\ell := t_\ell - t_{\ell-1} \le h_{max}, \quad h_{min} > 0, \quad \ell \ge 1,$$

$$\kappa_1 \le \frac{h_{\ell-1}}{h_\ell} \le \kappa_2, \quad \ell \ge 1,$$

where κ_1, κ_2, and h_{max} are suitable constants (cf. [14, 27]) such that there exists an $\mathbb{R}^k$-norm $\|\cdot\|_*$ with $\|\mathcal{F}_\ell\|_* \le 1$ for $\ell \ge k$ and all grids where

$$\mathcal{F}_\ell \;:=\; \begin{bmatrix} -\hat\alpha_{\ell,1} I & \cdots & \cdots & -\hat\alpha_{\ell,k} I \\ I & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & I & 0 \end{bmatrix},$$

and $\hat\alpha_{\ell,j}$ are related to the k-step BDF coefficients $\alpha_{\ell,j}$ by $\hat\alpha_{\ell,j} := \frac{\alpha_{\ell,j}}{\alpha_{\ell,0}}$, $j = 1, \ldots, k$, $k \le 6$. As in Section 6.1, we use the decoupling technique, stated in Chapter 5, to the BDF discretized equation (6.38). That is, we use the relations

$$\begin{aligned} A_\ell \;&=\; A_\ell D_\ell D_\ell^- \;=\; G_{2,\ell} P_{1,\ell} P_{0,\ell} D_\ell^- \\ B_\ell \;&=\; B_\ell \Pi_{1,\ell} + B_\ell \Pi_{0,\ell} Q_{1,\ell} + B_\ell Q_{0,\ell} \\ &=\; B_\ell \Pi_{1,\ell} + G_{2,\ell} P_{1,\ell} D_\ell^- \hat R_\ell' D_\ell \Pi_{0,\ell} Q_{1,\ell} + G_{2,\ell} Q_{1,\ell} + G_{2,\ell} Q_{0,\ell}, \end{aligned}$$

and transform the difference equation (6.38) into

$$\begin{aligned} G_{2,\ell} P_{1,\ell} P_{0,\ell} D_\ell^- [Dx]_\ell' + B_\ell \Pi_{1,\ell} x_\ell + G_{2,\ell} P_{1,\ell} D_\ell^- \hat R_\ell' D_\ell \Pi_{0,\ell} Q_{1,\ell} x_\ell \\ + G_{2,\ell} Q_{1,\ell} x_\ell + G_{2,\ell} Q_{0,\ell} x_\ell \;=\; G_{2,\ell} q_\ell, \end{aligned} \tag{6.41}$$

where $\hat R := D \Pi_1 D^-$ is introduced for shorter expressions.

Then, we scale equation (6.41) by $G_{2,\ell}^{-1}$ and multiply the resulting system by $D_\ell\Pi_{1,\ell}$, $\Pi_{0,\ell}Q_{1,\ell}$, and $Q_{0,\ell}P_{1,\ell}$, respectively. Further reformulations (cf. (5.7),(5.10),(5.9)) yield the decoupled system of (6.38)

$$\hat{R}_\ell[Dx]'_\ell + (D\Pi_1 G_2^{-1} B\Pi_1)_\ell x_\ell + \hat{R}'_\ell D_\ell \Pi_{0,li} Q_{1,li} x_\ell = (D\Pi_1 G_2^{-1} q)_\ell \tag{6.42a}$$

$$-(Q_0 Q_1 D^-)_\ell [Dx]'_\ell + (Q_0\Pi_1 G_2^{-1} B\Pi_1)_\ell x_\ell + Q_{0,li} x_\ell = (Q_0\Pi_1 G_2^{-1} q)_\ell \tag{6.42b}$$

$$\Pi_{0,\ell} Q_{1,li} x_\ell = (\Pi_0 Q_1 G_2^{-1} q)_\ell. \tag{6.42c}$$

Introducing the notation

$$[Mx]'_\ell \;:=\; \frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{\ell,j} M(t_{\ell-j}) x_{\ell-j},$$

for any matrix function $M(t)$ defined on $\mathcal{J}$ to derive

$$\begin{aligned}
[Dx]'_\ell \;&=\; \frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_{\ell-j} \\
&=\; \frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{\ell,j} (D_{\ell-j}\Pi_{0,\ell-j} P_{1,l-j} x_{\ell-j} + D_{\ell-j}\Pi_{0,\ell-j} Q_{1,l-j} x_{\ell-j}) \\
&=\; [D\Pi_1 x]'_\ell \;+\; [D\Pi_0 Q_1 x]'_\ell,
\end{aligned}$$

where we use the identities $D_{\ell-j} = D_{\ell-j}\Pi_{0,\ell-j}$ and $I = P_{1,\ell-j} + Q_{1,\ell-j}$.

Consequently, denoting $u_\ell := D_\ell\Pi_{1,\ell} x_\ell$ and $v_\ell := \Pi_{0,\ell} Q_{1,\ell} x_\ell$, the decoupled system (6.42) can be written as

$$\hat{R}_\ell[u]'_\ell + (D\Pi_1 G_2^{-1} BD^-)_\ell u_\ell + \hat{R}_\ell[Dv]'_\ell + \hat{R}'_\ell D_\ell v_\ell = (D\Pi_1 G_2^{-1} q)_\ell \tag{6.43a}$$

$$\begin{aligned}
&-(Q_0 Q_1 D^-)_\ell [Dv]'_\ell - (Q_0 Q_1 D^-)_\ell [u]'_\ell \\
&\qquad +(Q_0\Pi_1 G_2^{-1} BD^-)_\ell u_\ell + Q_{0,l} x_\ell = (Q_0\Pi_1 G_2^{-1} q)_\ell
\end{aligned} \tag{6.43b}$$

$$v_\ell = (\Pi_0 Q_1 G_2^{-1} q)_\ell. \tag{6.43c}$$

As in the case of Runge-Kutta method we consider the perturbed systems with the perturbations $\delta_\ell^{[1]}$ and $\delta_\ell^{[2]}$:

$$A_\ell[Dx^{[1]}]'_\ell + B_\ell x_\ell^{[1]} - q_\ell \;=\; \delta_\ell^{[1]}, \qquad \ell \ge k, \tag{6.44}$$

where $[Dx^{[1]}]'_\ell := \frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_{\ell-j}^{[1]}$ and

$$A_\ell[Dx^{[2]}]'_\ell + B_\ell x_\ell^{[2]} - q_\ell \;=\; \delta_\ell^{[2]}, \qquad \ell \ge k, \tag{6.45}$$

where $[Dx^{[2]}]'_\ell := \frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_{\ell-j}^{[2]}$.

Let $\Delta_\ell := \delta_\ell^{[1]} - \delta_\ell^{[2]}$. Subtracting the decoupled system of the difference equation (6.45) from the decoupled system of difference equation (6.44) we obtain

$$\begin{aligned}
&\hat{R}_\ell \left(u_\ell^{[1]} - u_\ell^{[2]}\right)' + (D\Pi_1 G_2^{-1} B D^-)_\ell (u_\ell^{[1]} - u_\ell^{[2]}) \\
&\quad + \hat{R}_\ell \left([Dv^{[1]}]'_\ell - [Dv^{[2]}]'_\ell\right) + \hat{R}'_\ell D_\ell (v_\ell^{[1]} - v_\ell^{[2]}) \qquad = \left(D\Pi_1 G_2^{-1}\Delta\right)_\ell \qquad (6.46a) \\
&- (Q_0 Q_1 D^-)_\ell \left([Dv^{[1]}]'_\ell - [Dv^{[2]}]'_\ell\right) - (Q_0 Q_1 D^-)_\ell \left(u_\ell^{[1]} - u_\ell^{[2]}\right)' \\
&\quad + (Q_0 \Pi_1 G_2^{-1} B D^-)_\ell (u_\ell^{[1]} - u_\ell^{[2]}) + Q_{0,\ell}(x_\ell^{[1]} - x_\ell^{[2]}) = \left(Q_0 P_1 G_2^{-1}\Delta\right)_\ell \qquad (6.46b) \\
&\qquad\qquad v_\ell^{[1]} - v_\ell^{[2]} = \left(\Pi_0 Q_1 G_2^{-1}\Delta\right)_\ell . \qquad (6.46c)
\end{aligned}$$

Theorem 6.6. *Let $x_* \in C_D^1([t_0, T], \mathbb{R}^m)$ be a solution of the index-2 DAE (6.1) with harmless critical points (cf. Definition 5.1). Let the partitions (6.2) with $\|\mathcal{F}_\ell\|_* \le 1$ for $\ell \ge k$ be given, and let the maximal stepsizes of all grids (6.2) be sufficiently small. Then, for all sufficiently small perturbations $|\delta_{\ell i}^{[1]}| \le \delta$, $|\delta_{\ell i}^{[2]}| \le \delta$, $\delta > 0$, and $x_\ell^{[1]}$ and $x_\ell^{[2]}$ related by (6.44), (6.45), respectively, the stability estimate*

$$\begin{aligned}
\|x_\ell^{[1]} - x_\ell^{[2]}\| \;\le\; K \Bigg(&\max_{j<k} \|D_j x_j^{[1]} - D_j x_j^{[2]}\| \;+\; \max_{k\le j\le \ell} \|\delta_j^{[1]} - \delta_j^{[2]}\| \\
&+ \max_{k\le j\le \ell} \left\{ \frac{1}{h_j} \|(D\Pi_0 Q_1 G_2^{-1})_j (\delta_j^{[1]} - \delta_j^{[2]})\| \right\} \Bigg) \qquad (6.47)
\end{aligned}$$

holds with a constant $K > 0$ independent of the stepsize.

Proof : Let $e_\ell := u_\ell^{[1]} - u_\ell^{[2]}$ and $f_\ell := v_\ell^{[1]} - v_\ell^{[2]}$ system (6.46) can be written as

$$\begin{aligned}
&\hat{R}_\ell e'_\ell + (D\Pi_1 G_2^{-1} B D^-)_\ell e_\ell + \hat{R}_\ell [Df]'_\ell + \hat{R}'_\ell D_\ell f_\ell = (D\Pi_1 G_2^{-1}\Delta)_\ell \qquad (6.48a) \\
&-(Q_0 Q_1 D^-)_\ell [Df]'_\ell - (Q_0 Q_1 D^-)_\ell e'_\ell \\
&\qquad +(Q_0 P_1 G_2^{-1} B D^-)_\ell e_\ell + Q_{0,\ell}(x_\ell^{[1]} - x_\ell^{[2]}) = (Q_0 P_1 G_2^{-1}\Delta)_\ell \qquad (6.48b) \\
&\qquad\qquad f_\ell = (\Pi_0 Q_1 G_2^{-1}\Delta)_\ell, \qquad (6.48c)
\end{aligned}$$

with $e'_\ell := \frac{1}{h_\ell} \sum_{j=0}^k \alpha_{\ell,j} e_{\ell-j}$ and $[Df]'_\ell := \frac{1}{h_\ell} \sum_{j=0}^k \alpha_{\ell,j} D_{\ell-j} f_{\ell-j}$.

Since $\hat{R}_\ell e_\ell = e_\ell$ and $\hat{R}_\ell D_\ell f_\ell = 0$ it follows that

$$\begin{aligned}
\hat{R}_\ell e'_\ell &= e'_\ell + \frac{1}{h_\ell} \sum_{j=0}^k \alpha_{\ell,j} \left(\hat{R}_\ell - \hat{R}_{\ell-j}\right) e_{\ell-j} \\
&= \frac{\alpha_{\ell,0}}{h_\ell} e_\ell + \frac{1}{h_\ell} \sum_{j=1}^k \alpha_{\ell,j} e_{\ell-j} + \frac{1}{h_\ell} \sum_{j=0}^k \alpha_{\ell,j} \left(\hat{R}_\ell - \hat{R}_{\ell-j}\right) e_{\ell-j}, \qquad (6.49) \\
\hat{R}_\ell [Df]'_\ell &= \frac{1}{h_\ell} \sum_{j=0}^k \alpha_{\ell,j} \left(\hat{R}_\ell - \hat{R}_{\ell-j}\right) D_{\ell-j} f_{\ell-j}. \qquad (6.50)
\end{aligned}$$

Then, inserting (6.48c), (6.49), and (6.50) into (6.48a) we obtain

$$e_\ell = -\sum_{j=1}^{k} \hat{\alpha}_{\ell,j} e_{\ell-j} + \frac{h_\ell}{\alpha_{\ell,0}} D_\ell \tilde{W}_\ell e_\ell + \frac{h_\ell}{\alpha_{\ell,0}} \varphi_\ell - h_\ell \phi_\ell - h_\ell \psi_\ell, \qquad \ell \geq k,$$

where the $\hat{\alpha}_{\ell,j}$ are related to the k-step BDF coefficients $\alpha_{\ell,j}$ by $\hat{\alpha}_{\ell,j} := \frac{\alpha_{\ell,j}}{\alpha_{\ell,0}}$ and

$$\begin{aligned}
\varphi_\ell &:= \left(D_\ell \Pi_{1,\ell} G_{2,\ell}^{-1} - \hat{R}'_\ell D_\ell \Pi_{0,\ell} Q_{1,\ell} G_{2,\ell}^{-1} \right) \Delta_\ell, \\
\psi_\ell &:= \frac{1}{h_\ell} \sum_{j=0}^{k} \hat{\alpha}_{\ell,j} (\hat{R}_\ell - \hat{R}_{\ell-j}) D_{\ell-j} f_{\ell-j}, \\
\phi_\ell &:= \frac{1}{h_\ell} \sum_{j=0}^{k} \hat{\alpha}_{\ell,j} (\hat{R}_\ell - \hat{R}_{\ell-j}) e_{\ell-j}, \\
\tilde{W}_\ell &:= -\Pi_{1,\ell} G_{2,\ell}^{-1} B_\ell D_\ell^- .
\end{aligned}$$

Thus,

$$\left(I - \frac{h_\ell}{\alpha_{\ell,0}} D_\ell \tilde{W}_\ell \right) e_\ell = -\sum_{j=1}^{k} \hat{\alpha}_{\ell,j} e_{\ell-j} + \frac{h_\ell}{\alpha_{\ell,0}} \varphi_\ell - h_\ell \phi_\ell - h_\ell \psi_\ell, \quad \ell \geq k. \tag{6.51}$$

Now, due to the continuity property of $D\Pi_1 G_2^{-1} B D^-$, there exists an h_* such that the matrix $\left(I - \frac{h_\ell}{\alpha_{\ell,0}} D_\ell \tilde{W}_\ell \right)$ is nonsingular for $h_\ell \leq h_*$ and

$$\left\| \left(I - \frac{h_\ell}{\alpha_{\ell,0}} D_\ell \tilde{W}_\ell \right)^{-1} \right\| \leq 1 + hC_1$$

for some constant C_1 independent of the stepsize. Denoting $H_\ell := \left(I - \frac{h_\ell}{\alpha_{\ell,0}} D_\ell \tilde{W}_\ell \right)$, we can write (6.51) as

$$\begin{aligned}
e_\ell &= -H_\ell^{-1} \sum_{j=1}^{k} \hat{\alpha}_{\ell,j} e_{\ell-j} + \frac{h_\ell}{\alpha_{\ell,0}} H_\ell^{-1} \varphi_\ell - h_\ell H_\ell^{-1} \phi_\ell - h_\ell H_\ell^{-1} \psi_\ell \\
&= -\left(I + \frac{h_\ell}{\alpha_{\ell,0}} D_\ell \tilde{W}_\ell H_\ell^{-1} \right) \sum_{j=1}^{k} \hat{\alpha}_{\ell,j} e_{\ell-j} + \frac{h_\ell}{\alpha_{\ell,0}} H_\ell^{-1} \left(\varphi_\ell - \alpha_{\ell,0} \phi_\ell - \alpha_{\ell,0} \psi_\ell \right),
\end{aligned} \tag{6.52}$$

where

$$\left(I - \frac{h_\ell}{\alpha_{\ell,0}} D_\ell \tilde{W}_\ell \right)^{-1} = \left(I + \frac{h_\ell}{\alpha_{\ell,0}} D_\ell \tilde{W}_\ell H_\ell^{-1} \right).$$

Together with the identities $e_\ell = e_\ell, \ldots, e_{\ell-k+1} = e_{\ell-k+1}$ we can rearrange the BDF recursion (6.52) to a one step recursion as

$$\underbrace{\begin{bmatrix} e_\ell \\ e_{\ell-1} \\ \vdots \\ e_{\ell-k+1} \end{bmatrix}}_{=:\mathcal{E}_\ell} = \underbrace{\begin{bmatrix} -\hat\alpha_{\ell,1}I & \cdots & \cdots & -\hat\alpha_{\ell,k}I \\ I & 0 & \cdots & 0 \\ & \ddots & \ddots & \vdots \\ 0 & & I & 0 \end{bmatrix}}_{=:\mathcal{F}_\ell} \underbrace{\begin{bmatrix} e_{\ell-1} \\ e_{\ell-2} \\ \vdots \\ e_{\ell-k} \end{bmatrix}}_{=:\mathcal{E}_{\ell-1}} + \frac{h_\ell}{\alpha_{\ell,0}} \underbrace{\begin{bmatrix} H_\ell^{-1}\varphi_\ell \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{=:\Omega_\ell}$$
$$- \frac{h_\ell}{\alpha_{\ell,0}} \underbrace{\begin{bmatrix} \alpha_{\ell,0}H_\ell^{-1}\phi_\ell \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{=:\Phi_\ell} - \frac{h_\ell}{\alpha_{\ell,0}} \underbrace{\begin{bmatrix} \alpha_{\ell,0}H_\ell^{-1}\psi_\ell \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{=:\Psi_\ell}$$
$$- \frac{h_\ell}{\alpha_{\ell,0}} \underbrace{\begin{bmatrix} D_\ell\tilde W_\ell H_\ell^{-1}\sum_{j=1}^k \hat\alpha_{\ell,j}e_{\ell-j} \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{=:\Xi_\ell},$$

or in compact form

$$\mathcal{E}_\ell = (\mathcal{F}_\ell \otimes I)\mathcal{E}_{\ell-1} + \frac{h_\ell}{\alpha_{\ell,0}}\Omega_\ell - \frac{h_\ell}{\alpha_{\ell,0}}\Phi_\ell - \frac{h_\ell}{\alpha_{\ell,0}}\Psi_\ell - \frac{h_\ell}{\alpha_{\ell,0}}\Xi_\ell.$$

Hence, taking into account that there is a norm such that $\|\mathcal{F}_\ell\|_* \le 1$ for all ℓ and letting $a_\ell = \frac{1}{|\alpha_{\ell,0}|}$, we have

$$\|\mathcal{E}_\ell\|_* \le \|\mathcal{E}_{\ell-1}\|_* + ha_\ell\|\Omega_\ell\|_* + ha_\ell\|\Phi_\ell\|_* + ha_\ell\|\Psi_\ell\|_* + ha_\ell\|\Xi_\ell\|_*.$$

Since $\|\Phi_\ell\|_* \le k_1\|\Delta_\ell\|$, $\|\Psi_\ell\|_* \le k_2\|\Delta_\ell\|$ and $\|\Xi_\ell + \Phi_\ell\|_* \le \tilde L\|\mathcal{E}_{\ell-1}\|_*$, using the standard recursion procedure we receive the required estimation for e_ℓ. Proceeding with $x_\ell^{[1]} - x_\ell^{[2]}$ as in the case of the Runge-Kutta method, we obtain the stability inequality. □

Now we can state the convergence result for k-step BDF ($k \le 6$) applied to index-2 DAE (6.1) with harmless critical points.

Theorem 6.7. *Let $x_* \in C_D^1([t_0,T],\mathbb{R}^m)$ be a solution of the index-2 DAE* (6.1) *with harmless critical points (cf. Definition 5.1). If the solution $x_*(\cdot)$ of the DAE* (6.1) *satisfies $D(\cdot)x_*(\cdot) \in C^{k+1}$ and the errors in the initial values have order $O(h^k)$, then the k-step BDF method* (6.38) *is convergent with order k.*

Proof : Consider the numerical solution x_ℓ generated by the k-step BDF ($k \le 6$)

$$A_\ell\frac{1}{h_\ell}\sum_{j=0}^k \alpha_{\ell,j}D_{\ell-j}x_{\ell-j} + B_\ell x_\ell \;=\; q_\ell, \qquad \ell \ge k.$$

Following [36], we define the local error τ_ℓ of the k-step BDF method applied to (6.1) as

$$\tau_\ell \;=\; A_\ell[Dx_*]'_\ell + B_\ell x_*(t_\ell) - q_\ell, \qquad \ell \geq k,$$

where

$$[Dx_*]'_\ell \;:=\; \frac{1}{h_\ell}\sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_{\ell-j}).$$

The local error τ_ℓ is accurate of order $O(h^k)$ if $D(\cdot)x_*(\cdot)$ is assumed to be sufficiently smooth. Further, τ_ℓ lies in $\operatorname{im} A_\ell$. This follows from the fact that $x_*(t_\ell)$ satisfies

$$A_\ell(Dx_*)'(t_\ell) + B_\ell x_*(t_\ell) - q_\ell \;=\; 0.$$

which make it possible to obtain

$$\begin{aligned}\tau_\ell \;&=\; A_\ell[Dx_*]'_\ell - A_\ell(Dx_*)'(t_\ell)\\ &=\; A_\ell\left(\frac{1}{h_\ell}\sum_{j=0}^{k}\alpha_{\ell,j}D_{\ell-j}x_*(t_{\ell-j}) - (Dx_*)'(t_\ell)\right).\end{aligned}$$

According to Theorem 6.6 we have

$$\begin{aligned}\|x_\ell - x_*(t_\ell)\| \;\leq\; K\Big(&\max_{j<k}\|D_j x_j - D_j x_*(t_j)\| \;+\; \max_{k\leq j\leq \ell}\|\tau_j\| \\ &+\; \max_{k\leq j\leq \ell}\left\{\frac{1}{h_j}\|(D\Pi_0 Q_1 G_2^{-1})_j \tau_j\|\right\}\Big).\end{aligned}$$

Due to $\tau_\ell \in \operatorname{im} A_\ell$, there exists a $w \in \mathbb{R}^m$ such that $\tau_\ell = A_\ell w$ is true. This implies $G_{2,\ell}^{-1}\tau_\ell = P_{1,\ell}P_{0,\ell}D_\ell^- w$ which leads to

$$D_\ell \Pi_{0,\ell} Q_{1,\ell} G_{2,\ell}^{-1}\tau_\ell \;=\; D_\ell \Pi_{0,\ell} Q_{1,\ell} P_{1,\ell} P_{0,\ell} D_\ell^- w \;=\; 0.$$

Therefore

$$\|x_\ell - x_*(t_\ell)\| \;\leq\; K\left(\max_{j<k}\|D_j x_j - D_j x_*(t_j)\| \;+\; \max_{k\leq j\leq \ell}\|\tau_j\|\right).$$

Finally, since $\tau_\ell = O(h^k)$ and $D_\ell x_\ell - D_\ell x_*(t_\ell) = O(h^k)$, $\ell = 0, \ldots, k-1$, the order of convergence is k. □

Chapter 7

Error estimation and stepsize prediction

For each numerical scheme that computes an approximate solution of a differential equation by a stepwise integration method, one has to decide whether to accept the results of a computed step or to repeat the step with a smaller stepsize. This decision is based on an estimate of the local error, i.e., the computed step is accepted if

$$\text{EST} \quad \leq \quad \text{TOL},$$

where EST is the local error estimate and TOL is the tolerance prescribed by the user. In this chapter we develop an estimator for the local error of numerical integration methods applied to linear index-2 DAEs which possess harmless critical points.

7.1 Error estimation and stepsize prediction for BDF methods

We first present an estimation of the local error and a stepsize selection algorithm for the BDF method,

$$A_\ell \frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_{\ell-j} + B_\ell x_\ell \;=\; q_\ell, \qquad \ell \geq k, \tag{7.1}$$

applied to linear index-2 DAE (6.1) with harmless critical points. Following [36], we define the *local error made in one step* by the BDF scheme (7.1) as the difference between the exact solution $x_*(t_\ell)$ and the numerical solution x_ℓ^* calculated from the values $x_*(t_{\ell-1})$, ..., $x_*(t_{\ell-k})$ replacing $x_{\ell-1}$, ..., $x_{\ell-k}$, respectively, in (7.1);

$$\Theta_\ell \;:=\; x_\ell^* - x_*(t_\ell), \qquad \ell \geq k. \tag{7.2}$$

Furthermore, we define the *local truncation error* at time point t_ℓ as

$$\lambda_\ell \;:=\; h_\ell \left((Dx_*)'(t_\ell) - \frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_{\ell-j}) \right), \qquad \ell \geq k. \tag{7.3}$$

Clearly, the local truncation error is defined as the error of the backward difference quotient (in Dx_*) scaled by h_ℓ. Recall from the relation (6.36) in Lemma 6.5 that if $(Dx_*)(\cdot) \in C^{(k+2)}[t_0, T]$ and a polynomial $p(t)$ interpolates the values $\{(t_{\ell-j}, (Dx_*)(t_{\ell-j})) : j = 0, 1, \ldots, k\}$, then the local truncation error λ_ℓ satisfies

$$\lambda_\ell \;=\; \frac{1}{(k+1)!} c_{\ell,k} (Dx_*)^{(k+1)}(t_\ell) h_\ell^{k+1} + O(h_\ell^{k+2}), \tag{7.4}$$

that is, the error λ_ℓ is of order $O(h_\ell^{k+1})$. The following result shows the relation between the errors Θ_ℓ and λ_ℓ.

Lemma 7.1. *Let the stepsize h_ℓ be sufficiently small to guarantee that the matrix $\left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right)$ is nonsingular. Then, the local errors Θ_ℓ and λ_ℓ satisfy the relation*

$$\Theta_\ell \;=\; \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right)^{-1} \left(\frac{1}{h_\ell} A_\ell \lambda_\ell\right), \qquad \ell \geq k, \tag{7.5}$$

where Θ_ℓ and λ_ℓ, for $\ell \geq k$, are specified in (7.2) *and* (7.3)*, respectively.*

Proof : From the definition (7.3) we can write $(Dx_*)'(t_\ell)$ as

$$(Dx_*)'(t_\ell) \;=\; \frac{1}{h_\ell}\lambda_\ell + \frac{\alpha_{\ell,0}}{h_\ell} D_\ell x_*(t_\ell) + \frac{1}{h_\ell} \sum_{j=1}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_\ell), \qquad \ell \geq k. \tag{7.6}$$

Since the exact solution $x_*(t_\ell)$ satisfies

$$A_\ell (Dx_*)'(t_\ell) + B_\ell x_*(t_\ell) \;=\; q_\ell, \tag{7.7}$$

inserting (7.6) into (7.7) leads to

$$\frac{1}{h_\ell} A_\ell \lambda_\ell \;=\; q_\ell - A_\ell \frac{1}{h_\ell} \sum_{j=1}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_{\ell-j}) - \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right) x_*(t_\ell).$$

Since x_ℓ^* fulfills the relation

$$\left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right) x_\ell^* \;=\; q_\ell - A_\ell \frac{1}{h_\ell} \sum_{j=1}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_{\ell-j}), \qquad \ell \geq k, \tag{7.8}$$

we obtain

$$\begin{aligned} \frac{1}{h_\ell} A_\ell \lambda_\ell \;&=\; \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right) x_\ell^* - \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right) x_*(t_\ell) \\ &=\; \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right) \left(x_\ell^* - x_*(t_\ell)\right). \end{aligned}$$

Therefore, using $x_\ell^* - x_*(t_\ell) = \Theta_\ell$, this statement is verified. □

Remark 7.2. *As in Chapter 6, if we define the local error as the defect obtained when the exact solution is inserted into the numerical scheme, i.e.*

$$\tau_\ell \;:=\; A_\ell \frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_{\ell-j}) + B_\ell x_*(t_\ell) - q_\ell, \qquad \ell \geq k, \tag{7.9}$$

we can find the relation between the local error Θ_ℓ and defect τ_ℓ, for $\ell \geq k$, as

$$\begin{aligned}
\tau_\ell \;&=\; \frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell x_*(t_\ell) + A_\ell \frac{1}{h_\ell} \sum_{j=1}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_{\ell-j}) + B_\ell x_*(t_\ell) - q_\ell \\
&=\; \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \right) x_*(t_\ell) - q_\ell + A_\ell \frac{1}{h_\ell} \sum_{j=1}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_{\ell-j}) \\
&=\; -\left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \right) (x_\ell^* - x_*(t_\ell)) \\
&=\; -\left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \right) \Theta_\ell.
\end{aligned}$$

If the matrix $\left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \right)$ is nonsingular, the following identity holds:

$$\Theta_\ell \;=\; -\left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \right)^{-1} \tau_\ell. \tag{7.10}$$

It is known that for higher indexes ($\mu > 1$) some components of the error are affected by the weak instability. In order to illustrate this influence we consider the regular index-2 DAE of the form

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} x(t) \right)' + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} x(t) \;=\; \begin{bmatrix} e^{-t}\sin(t) - 2\sin(2t) \\ \frac{1}{2}\left(2\cos t - 2\sin t + 8t - 5\right) e^{-t} - \cos(2t) + 1 \\ (t^2-1)\cos(2t) - 1 \\ \frac{1}{2}\left(2\sin t + 8t - 5\right) e^{-t} + (t^2-1)\cos(2t) \\ \cos(2t) - \sin t \cos t - 1 \end{bmatrix}. \tag{7.11}$$

This system is a regular variant of the DAE (3.10) introduced in Example 3.16 obtained by choosing $\alpha(t) = 1$, $\beta(t) = 1$ and $\gamma(t) = 1$. The exact solution is given by

$$\begin{aligned}
&x_1(t) = -\sin t \cos t, \quad x_2(t) = e^{-t} \sin t, \\
&x_3(t) = \cos(2t) - 1, \quad x_4(t) = \frac{1}{2}(8t - 5)\, e^{-t}, \quad x_5(t) = (t^2 - 1)\cos(2t).
\end{aligned}$$

We integrate the DAE (7.11) on the interval $[0, 2]$ by the order 2 BDF method (BDF_2) using different constant stepsizes h. Table 7.1 shows the order results for each component of the local error Θ_ℓ at $t = 2$. We observe that only the first three components of the local error Θ_ℓ have order 3, whereas the 4th and 5th components are of order 2. Recall from Example 3.16 that the first three components of the local error Θ_ℓ belong to the differential part of the system, whereas the 4th and 5th components are the algebraic Q_0-component of the system. This illustrates the fact that, for index-2 DAEs, only the algebraic part of the local error is amplified by $\frac{1}{h}$.

On the other hand, the following example shows that the weak instability does not affect the error arising when the integration methods apply to the index-1 systems.

Example 7.3. Let us consider a DAE

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & t^2+1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} x(t) \right)' + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} x(t) = q(t)$$

on the interval $\mathcal{J} = \mathbb{R}$. This DAE has a properly stated leading term and the product $G_0(t) = A(t)D$ reads

$$G_0(t) = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & t^2+1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Taking $R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $D^- = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ we obtain continuous projector functions $Q_0 = \mathrm{diag}(1, 0, 0, 1, 1)$, $P_0 = I - Q_0 = \mathrm{diag}(0, 1, 1, 0, 0)$ on $\mathcal{J}$. Thus,

$$G_1(t) = G_0(t) + BQ_0 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & t^2+1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

with $\det G_1(t) = t^2 + 1 \neq 0$. Therefore, the DAE is regular with index 1 on $\mathcal{J}$. The right-hand side q is given in such a way, that the exact solution is,

$$x_1(t) = \cos t - \frac{1}{2}e^{-3t}, \quad x_2(t) = -\sin t \cos t,$$

$$x_3(t) = 1 - e^{-3t}\sin t, \quad x_4(t) = 1 - t^2 \sin^2 t, \quad x_5(t) = \sin t \cos t - 1.$$

1st component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-01$	$1.0343e-03$		
$1.00e-02$	$1.3152e-06$	$2.8957e+00$	$8.1339e-01$
$1.00e-03$	$1.3424e-09$	$2.9911e+00$	$1.2625e+00$
$1.00e-04$	$1.3452e-12$	$2.9991e+00$	$1.3341e+00$
$1.00e-05$	$1.1102e-15$	$3.0834e+00$	$2.8993e+00$
$1.00e-06$	$5.5511e-17$		
2nd component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-01$	$1.7882e-03$		
$1.00e-02$	$1.8883e-06$	$2.9763e+00$	$1.6934e+00$
$1.00e-03$	$1.8960e-09$	$2.9983e+00$	$1.8732e+00$
$1.00e-04$	$1.8967e-12$	$2.9998e+00$	$1.8936e+00$
$1.00e-05$	$1.5682e-15$	$3.0826e+00$	$4.0593e+00$
$1.00e-06$	$1.3878e-16$		
3rd component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-01$	$1.0343e-03$		
$1.00e-02$	$1.3152e-06$	$2.8957e+00$	$8.1339e-01$
$1.00e-03$	$1.3424e-09$	$2.9911e+00$	$1.2625e+00$
$1.00e-04$	$1.3447e-12$	$2.9993e+00$	$1.3356e+00$
$1.00e-05$	$1.5543e-15$	$2.9371e+00$	$7.5332e-01$
$1.00e-06$	$0.0000e+00$		
4th component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-01$	$2.6366e-02$		
$1.00e-02$	$2.8651e-04$	$1.9639e+00$	$2.4263e+00$
$1.00e-03$	$2.8872e-06$	$1.9967e+00$	$2.8214e+00$
$1.00e-04$	$2.8895e-08$	$1.9997e+00$	$2.8805e+00$
$1.00e-05$	$2.3986e-10$	$2.0809e+00$	$6.0851e+00$
$1.00e-06$	$2.1764e-10$		
5th component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-01$	$2.4578e-02$		
$1.00e-02$	$2.8462e-04$	$1.9363e+00$	$2.1223e+00$
$1.00e-03$	$2.8853e-06$	$1.9941e+00$	$2.7696e+00$
$1.00e-04$	$2.8893e-08$	$1.9994e+00$	$2.8735e+00$
$1.00e-05$	$2.3986e-10$	$2.0808e+00$	$6.0832e+00$
$1.00e-06$	$2.1764e-10$		

Table 7.1: Observed order of the local error Θ_ℓ for the order 2 BDF of the problem (7.11) with different constant stepsizes. The numerically observed order is 3 in the first three components. For the 4th and 5th components the methods show order 2 behavior of the local error Θ_ℓ.

Consequently, the system of DAE reads

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_3(t) \\ x_2(t) \\ 0 \end{bmatrix}' + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} x(t) \;=\; \begin{bmatrix} e^{-3t}(3\sin t - \cos t) - \sin t \cos t \\ \sin^2 t - (t^2+1)\cos^2 t + e^{-3t}\sin t \\ 1 - e^{-3t}\sin t + \sin t \cos t - 1 \\ -t^2 \sin^2 t \\ e^{-3t}\left(-\sin t - \frac{1}{2}\right) + \cos t + 1 \end{bmatrix}. \tag{7.12}$$

As for the DAE (7.11), we apply the order 2 BDF to the problem (7.12) on the interval $[0, 2]$ with different values of the stepsize h. The order results of the local error Θ_ℓ at $t = 2$ are shown in Table 7.2. This indicates that for the index-1 DAEs all components of the local error Θ_ℓ have order 3 and no influence of the weak instability is visible.

Error estimation

In this part we provide an estimate of the local truncation error λ_ℓ for the BDF method (7.1). This approach has been developed in [17] to approximate the local truncation error (in x') for a BDF method applied to approximate solutions of the DAEs (6.8) and was used in [89, 91] for the error control of the BDF scheme applied to the quasi-linear index-2 problem (6.40). Furthermore, it has been extended in [20] to estimate the local truncation error (in d') for the BDF method applied to properly stated index-1 DAEs of the form $A\,(x(t), t)\,(d(x(t), t))' + b\,(x(t), t) = 0$.

Assume that the analytical solution $x_*(\cdot)$ is in $C^{(k+2)}([t_0, T], \mathbb{R}^n)$. The local truncation error λ_ℓ of the BDF scheme (7.1) can be estimated as follows (cf. [6, 17, 20]): For $\ell \geq k+1$, let $p_{\ell,k+1}(t)$ be the unique polynomial of degree $k+1$ which interpolates the values $\{(t_{\ell-j}, (Dx_*)(t_{\ell-j})) : j = 0, 1, \ldots, k+1\}$. Then we have

$$(Dx_*)'(t) - (p_{\ell,k+1}(t))' = O(h_\ell^{k+1}), \qquad t \in [t_{\ell-k-1}, t_\ell].$$

Further, let $p_{\ell,k}(t)$ be the unique polynomial of degree k which interpolates the values $\{(t_{\ell-j}, (Dx_*)(t_{\ell-j}) : j = 0, 1, \ldots, k\}$. By the construction of the BDF method (see [46, 85]) it follows that

$$(Dx_*)'(t_\ell) + O(h_\ell^k) \;=\; (p_{\ell,k}(t_\ell))' \;=\; \frac{1}{h_\ell} \sum_{j=0}^{k} \alpha_{\ell,j} D_{\ell-j} x_*(t_\ell).$$

Hence, λ_ℓ can be written as

$$\lambda_\ell \;=\; h_\ell \left\{(p_{\ell,k+1}(t_\ell))' - (p_{\ell,k}(t_\ell))'\right\} + O(h_\ell^{k+2}).$$

1st component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-001$	$6.340e-005$		
$1.00e-002$	$1.919e-008$	$3.519e+000$	$2.094e-001$
$1.00e-003$	$1.539e-011$	$3.096e+000$	$2.985e-002$
$1.00e-004$	$1.499e-014$	$3.012e+000$	$1.666e-002$
$1.00e-005$	$1.110e-016$	$2.130e+000$	$4.978e-006$
$1.00e-006$	$1.110e-016$		
2nd component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-001$	$6.822e-004$		
$1.00e-002$	$5.918e-007$	$3.062e+000$	$7.865e-001$
$1.00e-003$	$5.821e-010$	$3.007e+000$	$6.117e-001$
$1.00e-004$	$5.812e-013$	$3.001e+000$	$5.848e-001$
$1.00e-005$	$4.996e-016$	$3.066e+000$	$1.064e+000$
$1.00e-006$	$5.551e-017$		
3rd component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-001$	$6.340e-005$		
$1.00e-002$	$1.919e-008$	$3.519e+000$	$2.094e-001$
$1.00e-003$	$1.539e-011$	$3.096e+000$	$2.985e-002$
$1.00e-004$	$1.499e-014$	$3.012e+000$	$1.666e-002$
$1.00e-005$	$1.110e-016$	$2.130e+000$	$4.978e-006$
$1.00e-006$	$1.110e-016$		
4th component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-001$	$7.456e-004$		
$1.00e-002$	$6.110e-007$	$3.086e+000$	$9.100e-001$
$1.00e-003$	$5.975e-010$	$3.010e+000$	$6.390e-001$
$1.00e-004$	$5.960e-013$	$3.001e+000$	$6.021e-001$
$1.00e-005$	$4.441e-016$	$3.128e+000$	$1.933e+000$
$1.00e-006$	$4.441e-016$		
5th component			
h	$\|\Theta_\ell\| = \|x^*_\ell - x_*(t_\ell)\|$	**order**	**const.**
$1.00e-001$	$6.340e-005$		
$1.00e-002$	$1.919e-008$	$3.519e+000$	$2.094e-001$
$1.00e-003$	$1.539e-011$	$3.096e+000$	$2.985e-002$
$1.00e-004$	$1.488e-014$	$3.015e+000$	$1.704e-002$
$1.00e-005$	$2.220e-016$	$1.826e+000$	$2.998e-007$
$1.00e-006$	$0.000e+000$		

Table 7.2: Observed order of the local error Θ_ℓ for the order 2 BDF of the problem (7.12) with different constant stepsizes. The numerically observed order is 3 for all components of the local error Θ_ℓ.

The polynomials $p_{\ell,k+1}(t)$ and $p_{\ell,k}(t)$ can be written in Newton form [88]:

$$\begin{aligned} p_{\ell,k}(t) &= a_0 + a_1(t-t_\ell) + \cdots + a_k(t-t_\ell)\cdots(t-t_{\ell-k+1}), \\ p_{\ell,k+1}(t) &= p_{\ell,k}(t) + a_{k+1}(t-t_\ell)\cdots(t-t_{\ell-k}). \end{aligned}$$

Therefore,

$$\begin{aligned} \lambda_\ell &= h_\ell a_{k+1} \frac{d}{dt}(t-t_\ell)\cdots(t-t_{\ell-k}) \mid_{t=t_\ell} + O(h_\ell^{k+2}) \\ &= h_\ell a_{k+1} \prod_{i=1}^{k}(t_\ell - t_{\ell-i}) + O(h_\ell^{k+2}), \end{aligned}$$

As shown in [17, 20], the representation of the coefficient a_{k+1} can be computed from the comparison of Newton and Lagrange forms,

$$a_{k+1} = \sum_{j=0}^{k+1} \frac{1}{\prod_{i=0,i\neq j}^{k+1}(t_{\ell-j} - t_{\ell-i})} D_{\ell-j}x_*(t_{\ell-j}),$$

which allows to formulate λ_ℓ as

$$\lambda_\ell = h_\ell \frac{\prod_{i=1}^{k}(t_\ell - t_{\ell-i})}{\prod_{i=1}^{k+1}(t_\ell - t_{\ell-i})} \left[D_\ell x_*(t_\ell) + \sum_{j=1}^{k+1} \frac{\prod_{i=1}^{k+1}(t_\ell - t_{\ell-i})}{\prod_{i=0,i\neq j}^{k+1}(t_{\ell-j} - t_{\ell-i})} D_{\ell-j}x_*(t_{\ell-j}) \right] + O(h_\ell^{k+2}). \quad (7.13)$$

Let $\tilde{p}(t)$ be the unique polynomial of degree k that interpolates $\{(t_{\ell-j}, (Dx_*)(t_{\ell-j})) : j = 1, 2, \ldots, k+1\}$. Then,

$$\begin{aligned} \tilde{p}(t_\ell) &= \sum_{j=1}^{k+1} \prod_{i=1,i\neq j}^{k+1} \frac{t_\ell - t_{\ell-i}}{t_{\ell-j} - t_{\ell-i}} D_{\ell-j}x_*(t_{\ell-j}) \\ &= \sum_{j=1}^{k+1} \frac{\prod_{i=1}^{k+1}(t_\ell - t_{\ell-i})}{-(t_{\ell-j} - t_\ell)\prod_{i=1,i\neq j}^{k+1}(t_{\ell-j} - t_{\ell-i})} D_{\ell-j}x_*(t_{\ell-j}) \\ &= -\sum_{j=1}^{k+1} \frac{\prod_{i=1}^{k+1}(t_\ell - t_{\ell-i})}{\prod_{i=0,i\neq j}^{k+1}(t_{\ell-j} - t_{\ell-i})} D_{\ell-j}x_*(t_{\ell-j}). \end{aligned}$$

Inserting this equation into (7.13) gives

$$\lambda_\ell = \frac{h_\ell}{t_\ell - t_{\ell-k-1}} (D_\ell x_*(t_\ell) - \tilde{p}(t_\ell)) + O(h_\ell^{k+2}). \quad (7.14)$$

In a numerical computation, the exact solutions $x_*(t_{\ell-j})$, $j = 0, 1, \ldots, k+1$, are not available, but we have the approximations $x_{\ell-j}$. By the application of an order k method the relation $x_{\ell-j} = x_*(t_{\ell-j}) + O(h_\ell^{k+1})$ holds. Moreover, the polynomial

$\tilde{p}(t)$ interpolates also the value $\{(t_{\ell-j}, D_{\ell-j}x_{\ell-j}) : j = 1, 2, \ldots, k+1\}$. Therefore, the local truncation error λ_ℓ can be estimated by

$$\hat{\lambda}_\ell \;=\; \frac{h_\ell}{t_\ell - t_{\ell-k-1}} \left(D_\ell x_\ell - D_\ell x_\ell^p\right), \qquad \ell \geq k, \tag{7.15}$$

where $D_\ell x_\ell^p := \tilde{p}(t_\ell)$ and terms of order $O(h_\ell^{k+2})$ have been neglected.

Unfortunately, the local error estimate $\hat{\lambda}_\ell$ does not posses the asymptotic correct property in the index-2 case, as stated, e.g., in [82, 83] for the linear standard form DAEs with constant coefficients. However, according to the identity (5.17) in Theorem 5.2 and the relationship (7.5) between the errors Θ_ℓ and λ_ℓ in Lemma 7.1, we may develop an appropriate local error estimate for BDF methods applied to linear index-2 DAE which holds critical points

Using the relation (7.5) and the estimate (7.15), we can estimate Θ_ℓ by

$$\hat{\Theta}_\ell \;=\; \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right)^{-1} \frac{1}{h_\ell} A_\ell \hat{\lambda}_\ell, \tag{7.16}$$

and by the identity (5.17) in Theorem 5.2, the error estimate $\hat{\Theta}_\ell$ may be written as

$$\begin{aligned}\hat{\Theta}_\ell = \Big\{ &-\frac{1}{h_\ell} Q_0 Q_1 + \frac{1}{\alpha_{\ell,0}} \Pi_{\text{can2}} + \Pi_{\text{can2}} \cdot O(h_\ell) \cdot \Pi_1 + \frac{1}{\alpha_{\ell,0}} Q_0 Q_1 D^-(D\Pi_1 D^-)' D\Pi_1 \\ &+ Q_0 Q_1 D^-(D\Pi_1 D^-)' D\Pi_1 \cdot O(h_\ell) \cdot \Pi_1 \Big\} D_\ell^- \hat{\lambda}_\ell,\end{aligned} \tag{7.17}$$

where the property $A = ADD^-$ is used. As in the case of Θ_ℓ, this expression shows that the local error estimate $\hat{\Theta}_\ell$ does not behave as $O(h_\ell^{k+1})$ in some components, although the error estimate $\hat{\lambda}_\ell$ does. In [82, 83] it has been shown that the estimate $\hat{\Theta}_\ell$ is not a reliable enough for the error control of the BDF method applied to linear standard form index-2 DAEs with constant coefficients. As we have seen in (7.17) the error estimate $\hat{\Theta}_\ell$ inherits a part multiplied by $\frac{1}{h_\ell} Q_0 Q_1$. One may, however, eliminate this term by applying the identity (5.27) from Section 5.1 in Chapter 5 to derive an error estimate of order $O(h_\ell^{k+1})$. This technique has been proposed in [82, 83] to approximate the local error for the BDF scheme of linear standard form DAEs with constant coefficients. This error estimate has also been used in [38] for the implementation of a BDF scheme to solve the DAE systems arising from the method of lines discretization up to and including index-2 problems.

Motivated by (5.27), the error estimate for the BDF method (7.1) reads

$$\hat{\nu}_\ell \;=\; \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right)^{-1} \frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell\right)^{-1} \frac{1}{h_\ell} A_\ell \hat{\lambda}_\ell, \tag{7.18}$$

where $\hat{\lambda}_\ell$ is calculated from (7.15). Furthermore, it follows from (5.27) that

$$\hat{\nu}_\ell \;=\; \Big\{ \frac{1}{\alpha_{\ell,0}} \Pi_{\text{can2}} + \Pi_{\text{can2}} \cdot O(h_\ell) \cdot \Pi_1 + \frac{1}{\alpha_{\ell,0}} Q_0 Q_1 D^- (D\Pi_1 D^-)' D\Pi_1 \\ + Q_0 Q_1 D^- (D\Pi_1 D^-)' D\Pi_1 \cdot O(h_\ell) \cdot \Pi_1 \Big\} D_\ell^- \hat{\lambda}_\ell \tag{7.19}$$

which implies that the components multiplied by Q_0Q_1 are still included in the local error estimate $\hat{\nu}_\ell$. In order to exclude these components from the estimate we may scale $\hat{\nu}_\ell$ by the coefficient matrix D_ℓ, i.e., we use the property $D_\ell Q_{0,\ell} = 0$. Consequently, scaling (7.18) by D_ℓ yields

$$D_\ell \hat{\nu}_\ell \;=\; D_\ell \Big(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \Big)^{-1} \frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell \Big(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \Big)^{-1} \frac{1}{h_\ell} A_\ell \hat{\lambda}_\ell, \tag{7.20}$$

and simplifies the expression (7.19) to

$$D_\ell \hat{\nu}_\ell \;=\; \Big\{ \frac{1}{\alpha_{\ell,0}} D_\ell \Pi_{\text{can2}} + D_\ell \Pi_{\text{can2}} \cdot O(h_\ell) \cdot \Pi_1 \Big\} D_\ell^- \hat{\lambda}_\ell. \tag{7.21}$$

Due to equation (7.17), when we scale the estimate $\hat{\Theta}_\ell$ in (7.16) by D_ℓ, we obtain

$$D_\ell \hat{\Theta}_\ell \;=\; \Big\{ \frac{1}{\alpha_{\ell,0}} D_\ell \Pi_{\text{can2}} + D_\ell \Pi_{\text{can2}} \cdot O(h_\ell) \cdot \Pi_1 \Big\} D_\ell^- \hat{\lambda}_\ell. \tag{7.22}$$

The approaches used to avoid problems which are due to the instable Q_0Q_1-part, have been studied in [89, 91]. There, it has been stated that the numerical integration will work better than when the stepsize control is based on the differential components only. These components are not affected by the weak instability.

In the following example we compare the behavior of the error estimations $\hat{\lambda}_\ell$, $\hat{\nu}_\ell$, and $\hat{\Theta}_\ell$ for the BDF method applied to index-2 problem with harmless critical points.

Example 7.4. Index-2 DAE with harmless critical points ($m = 5, n = 3$)

Consider the DAE (3.16), introduced in Section 3.2.2, with

$$A(t) = \begin{bmatrix} \alpha(t) & 0 & 0 \\ 0 & \beta(t) & 0 \\ 0 & 0 & \gamma(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix},$$

where $\alpha(t)$, $\beta(t)$, and $\gamma(t)$ are continuous functions and $t \in \mathcal{J} = \mathbb{R}$. We choose $\alpha = 1$, $\beta = 1$ and $\gamma(t) = \sin(4t)$, set

$$x_1(t) = -\sin t \cos t, \quad x_2(t) = e^{-t} \sin t,$$
$$x_3(t) = \cos(2t) - 1, \quad x_4(t) = \frac{1}{2}(8t - 5)\, e^{-t}, \quad x_5(t) = (t^2 - 1)\cos(2t),$$

and then compute q. Consequently, the system of DAE reads

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sin(4t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_3(t) \\ x_2(t) \\ x_1(t) \end{bmatrix}' + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} x(t) \; = \;$$

$$\begin{bmatrix} e^{-t}\sin(t) - 2\sin(2t) \\ \frac{1}{2}\left(2\cos t - 2\sin t + 8t - 5\right)e^{-t} - \cos(2t) + 1 \\ (t^2 - \sin(4t))\cos(2t) - 1 \\ \frac{1}{2}\left(2\sin t + 8t - 5\right)e^{-t} + (t^2 - 1)\cos(2t) \\ \cos(2t) - \sin t \cos t - 1 \end{bmatrix}. \tag{7.23}$$

As described in Example 3.16 the zeros of $\gamma(t)$ define harmless critical points. Hence, this DAE has harmless critical points at $t = \pm\frac{n\pi}{4}$, $n = 0, 1, \ldots$. We applied the order 2 BDF method to the system (7.23) using constant stepsize $h = 0.1$. Figure 7.1 displays the behavior of the error estimates for the differential components, i.e. for the first three components of the error estimations. Thereby, EST(λ_l), EST(ν_l), and EST(Θ_l) correspond to the error estimates $\hat{\lambda}_\ell$, $\hat{\nu}_\ell$, and $\hat{\Theta}_\ell$, respectively. The results show that the error estimate $\hat{\Theta}_\ell$ provides a better basis for the error control. As already mentioned, the differential components of the error estimations $\hat{\Theta}_\ell$ and $\hat{\nu}_\ell$ are in fact $D_\ell\hat{\Theta}_\ell$ and $D_\ell\hat{\nu}_\ell$, respectively.

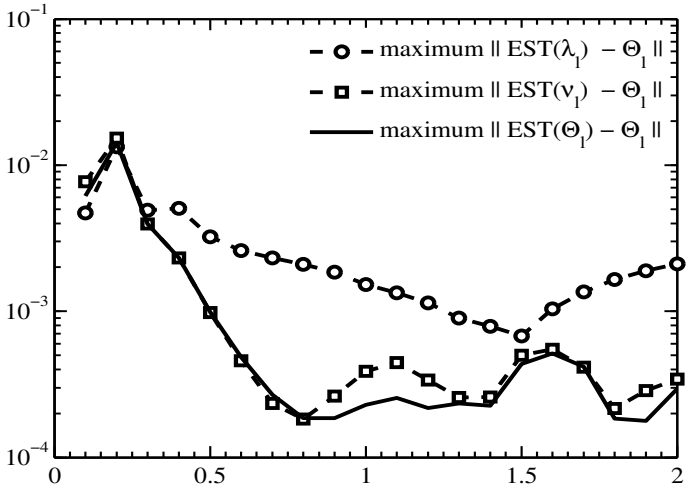


Figure 7.1: Comparison of the maximum norm $\|\hat{\lambda}_\ell - \Theta_\ell\|_\infty$ (dashed line with $\circ$), $\|\hat{\nu}_\ell - \Theta_\ell\|_\infty$ (dash dotted line with $\square$), and $\|\hat{\Theta}_\ell - \Theta_\ell\|_\infty$ (solid line), respectively, for the differential components of the error estimations of the DAE (7.23). All calculations were carried out by the BDF_2 with fixed stepsize $h = 0.1$ on the interval $[0, 2]$.

Since the estimate $\hat{\Theta}_\ell$ requires the factorization of the matrix $\left(\frac{\alpha_{\ell,0}}{h_\ell}A_\ell D_\ell + B_\ell\right)$ less than the estimate $\hat{\nu}_\ell$, we will use the estimate $\hat{\Theta}_\ell$ in the error control and the

stepsize selection algorithm. Therefore, the local error estimate is given by

$$D_\ell \hat{\Theta}_\ell \;=\; D_\ell \left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \right)^{-1} \frac{1}{h_\ell} A_\ell \hat{\lambda}_\ell, \tag{7.24}$$

where $D_\ell = D(t_\ell)$ and $\hat{\lambda}_\ell$ is computed from (7.15).

The proposed error estimate $D_\ell \hat{\Theta}_\ell$ is also valid for DAEs with lower index (1 or 0). In these cases we have the projectors $Q_1 = 0$ (index 1) and $Q_0 = 0$ (index 0), due to the nonsingularity of the matrices G_1 or G_0. The canonical projector Π_{can2} is then simplified to

$$\Pi_{\text{can2}} \;=\; \left(I - Q_0 G_1^{-1} B \Pi_0 \right) \Pi_0 \;=\; \Pi_{\text{can1}}$$

for index-1 case and to $\Pi_{\text{can2}} = I$ for index-0 problem or ODE. Therefore, the expression of the error estimate reads

$$D_\ell \hat{\Theta}_\ell \;=\; \left\{ \frac{1}{\alpha_{\ell,0}} D_\ell \Pi_{\text{can1}} + D_\ell \Pi_{\text{can1}} \cdot O(h_\ell) \cdot \Pi_0 \right\} D_\ell^- \hat{\lambda}_\ell,$$

for index-1 DAEs and

$$D_\ell \hat{\Theta}_\ell \;=\; \hat{\Theta}_\ell \;=\; \left\{ \frac{1}{\alpha_{\ell,0}} \cdot I + O(h_\ell) \right\} \hat{\lambda}_\ell,$$

for ODEs.

Remark 7.5. *The realization of this error estimate requires the solution of a linear system of equations, i.e., the matrix* $\left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \right)$ *and its factorization are needed. Since the Jacobian of the BDF method* (7.1) *approximates this matrix, we can use the LU decompositions to realize the matrix* $\left(\frac{\alpha_{\ell,0}}{h_\ell} A_\ell D_\ell + B_\ell \right)$*. Therefore, the error estimate based on the relation* (7.5) *does not cost additional evaluations.*

Stepsize prediction

In order to decide whether to accept or reject the results of the current step we define

$$e_\ell \;:=\; \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\left| D_{\ell,i} \hat{\Theta}_{\ell,i} \right|}{\texttt{tol}_i} \right)^2 \right)^{\frac{1}{2}}, \tag{7.25}$$

where $\left| D_{\ell,i} \hat{\Theta}_{\ell,i} \right| = \left| D(t_{\ell,i}) \hat{\Theta}_{\ell,i} \right|$ is calculated according to (7.24) and $\texttt{tol}_i$ is computed componentwise as

$$\texttt{tol}_i \;=\; \texttt{atol}_i + \texttt{rtol}_i \cdot \left| D_{\ell,i} x_{\ell,i} \right|,$$

with the error tolerances `atol` and `rtol` prescribed by the user. The indices i correspond to the i-th component of the vector. If $e_\ell \leq 1$, the computed step is accepted and the next integration step will be carried out with $h_{\ell+1} = h_{new}$. If

$e_\ell > 1$, the step is rejected and the computations are repeated with the smaller stepsize, i.e., with $h_{\ell,new} = h_{new}$.

Once the current step has been accepted or rejected, the new stepsize for the next step or for the repeat step have to be chosen as follows: For the method of order k a stepsize selection bases on the assumption that the local error Θ_ℓ has a representation (cf. (7.4) and (7.5))

$$\Theta_\ell \;=\; \phi_\ell h_\ell^{k+1} + O(h_\ell^{k+2}),$$

with a slowly varying, h-independent function ϕ. If a local error Θ_ℓ is observed for a given stepsize h_ℓ, then a local error $\Theta_{\ell+1}$ is obtained in time step $t_{\ell+1}$ as

$$\Theta_{\ell+1} \;=\; \phi_{\ell+1} h_{new}^{k+1} + O(h_\ell^{k+2}).$$

Since optimal performance is realized if the local error coincides with the desired tolerance, we aim to choose h_{new} such that $|D_{\ell+1}\Theta_{\ell+1}| \approx \texttt{tol}$. This can be achieved by requiring

$$\left(\frac{h_{new}}{h_\ell}\right)^{k+1} \;\approx\; \frac{\texttt{tol}}{|D_\ell \Theta_\ell|}$$

assumed that $\phi_\ell = \phi_{\ell+1}$ and neglecting the higher order term. Since $|D_\ell \hat{\Theta}_\ell| \approx |D_\ell \Theta_\ell|$,

$$\left(\frac{h_{new}}{h_\ell}\right)^{k+1} \;\approx\; \frac{\texttt{tol}}{\left|D_\ell \hat{\Theta}_\ell\right|}$$

or, equivalently,

$$h_{new} \;\approx\; \left(\frac{\texttt{tol}}{\left|D_\ell \hat{\Theta}_\ell\right|}\right)^{\frac{1}{k+1}} \cdot h_\ell.$$

As usual, a safety factor fac, say $fac = 0.7$, will be employed to avoid undesirable oscillatory behavior of the stepsize sequence which might destroy stability properties of the methods [33]. Hence, we let

$$h_{new} \;:=\; fac \cdot (e_\ell)^{-\frac{1}{k+1}} \cdot h_\ell. \tag{7.26}$$

This is the elementary stepsize control. More stepsize control strategies can be constructed as in [31, 86, 87]. For instance, the new stepsize may be computed according to the predictive stepsize control proposed by Gustafsson for implicit RK methods in [30] and investigated by Sjö [85] for BDF methods:

$$\begin{aligned} h_{new} \;&=\; fac \cdot \left(\frac{\texttt{tol}}{|D_\ell \hat{\Theta}_\ell|}\right)^{\frac{1}{k+1}} \cdot \left(\frac{|D_{\ell-1}\hat{\Theta}_{\ell-1}|}{|D_\ell \hat{\Theta}_\ell|}\right)^{\frac{1}{k+1}} \cdot \left(\frac{h_\ell}{h_{\ell-1}}\right) \cdot h_\ell \\ &=\; fac \cdot (e_\ell)^{-\frac{2}{k+1}} \cdot (e_{\ell-1})^{\frac{1}{k+1}} \cdot (h_\ell)^2 \cdot (h_{\ell-1})^{-1}, \end{aligned} \tag{7.27}$$

or one may apply the proportional integral control (PI control) [31] which has the form

$$h_{new} \;=\; fac \cdot (e_\ell)^{-\frac{0.7}{k+1}} \cdot (e_{\ell-1})^{\frac{0.4}{k+1}} \cdot h_\ell. \tag{7.28}$$

Note that in case of step rejections the control should be restarted by applying the elementary control (7.26).

Numerical results

The approach described in the previous part has been implemented for the BDF method with controlled stepsize using MATLAB. This solver is a variable coefficient implementation of a constant order BDF method which is capable of solving linear index-2 problems, particularly, for solving the index-2 DAEs with harmless critical points (cf. Chapter 5). In all tested examples we illustrated the performance of the error control using the estimators $\hat{\lambda}_\ell$ in (7.15), $D_\ell\hat{\nu}_\ell$ in (7.20), and $D_\ell\hat{\Theta}_\ell$ in (7.24). We plotted the global error of the numerical solution and the number of integration steps with respect to the different tolerances `rtol` in logarithmic scale. In all cases we used the elementary stepsize control (7.26) with $fac = 0.7$.

Example 7.6. Stiff problem

We start with a simple tested problem introduced in [74]. Consider the scalar initial value problem

$$x'(t) = L(x(t) - g(t)) + g'(t), \qquad x(0) = g(0), \qquad t \in [0, 10], \tag{7.29}$$

where $g(t) = \sin(t)$. The exact solution is given by the smooth function $x(t) = g(t)$. If L is negative and large in magnitude, then the problem can be arbitrarily stiff. For this example, assume $L = -100$.

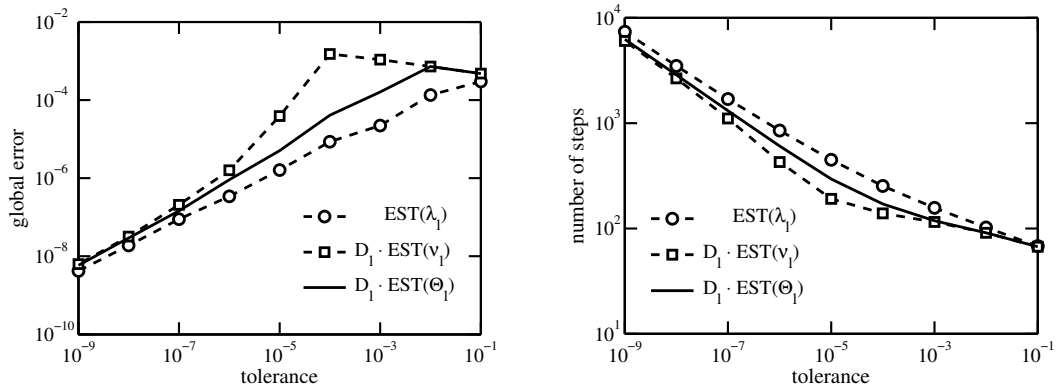


Figure 7.2: Stiff problem: The global error at the end of the integration interval (left) and the number of steps (right) generated by the BDF_2 according to $\hat{\lambda}_\ell$ in (7.15), $D_\ell\hat{\nu}_\ell$ in (7.20), and $D_\ell\hat{\Theta}_\ell$ in (7.24), respectively, vs. the used tolerances. Lines with $\mathrm{EST}(\lambda_l)$, $\mathrm{D_l}\cdot\,\mathrm{EST}(\nu_l)$, and $\mathrm{D_l}\cdot\,\mathrm{EST}(\Theta_l)$ correspond to the error estimations $\hat{\lambda}_\ell$, $D_\ell\hat{\nu}_\ell$, and $D_\ell\hat{\Theta}_\ell$, respectively.

We integrated this problem with the BDF_2 method on the interval $[0, 10]$ with the

relative tolerances $\texttt{rtol} = 10^{-j}$, $j = 1, \ldots, 9$. The absolute tolerances satisfied $\texttt{atol} = \texttt{rtol}$. The initial stepsize was chosen as $h_1 = 10^{-2} \cdot (\texttt{rtol} + \texttt{atol})$. Figure 7.2 presented the performances of the error estimates generated by the BDF_2 using the estimates $\hat{\lambda}_\ell$, $D_\ell \hat{\nu}_\ell$, and $\hat{\Theta}_\ell$, respectively. Thereby, lines with $\text{EST}(\lambda_1)$, $\text{D}_1 \cdot \text{EST}(\nu_1)$, and $\text{D}_1 \cdot \text{EST}(\Theta_1)$ correspond to the error estimates $\hat{\lambda}_\ell$, $D_\ell \hat{\nu}_\ell$, and $D_\ell \hat{\Theta}_\ell$, respectively. In the left graph of Figure 7.2 we depicted the global error at the end of the integration interval with respect to the applied tolerances `rtol`, while the number of steps vs. the used tolerances `rtol` was plotted in the right of diagram, both in logarithmic scale. The results indicate that using the estimator $\hat{\lambda}_\ell$ the numerical integration method show better performances than the ones with $D_\ell \hat{\nu}_\ell$ and $D_\ell \hat{\Theta}_\ell$ but require fewer integration steps.

Example 7.7. Regular index-1 DAE

We reconsider the linear index-1 DAE (7.12) considered in Example 7.3. We solved this problem with the BDF_2 method on the interval $[0, 2]$ using constant stepsize $h = 0.1$. The maximum norms of the difference between the exact local error Θ_ℓ and the local error estimates $\hat{\lambda}_\ell$, $D_\ell \hat{\nu}_\ell$, and, $D_\ell \hat{\Theta}_\ell$ specified in (7.15), (7.20), (7.24), respectively, were displayed in Figure 7.3. The results show that the error estimate $D_\ell \hat{\Theta}_\ell$ can approximate Θ_ℓ better than the error estimates $\hat{\lambda}_\ell$ and $D_\ell \hat{\nu}_\ell$. Only at some points the estimators $\hat{\lambda}_\ell$ or $D_\ell \hat{\nu}_\ell$ give a better performance.

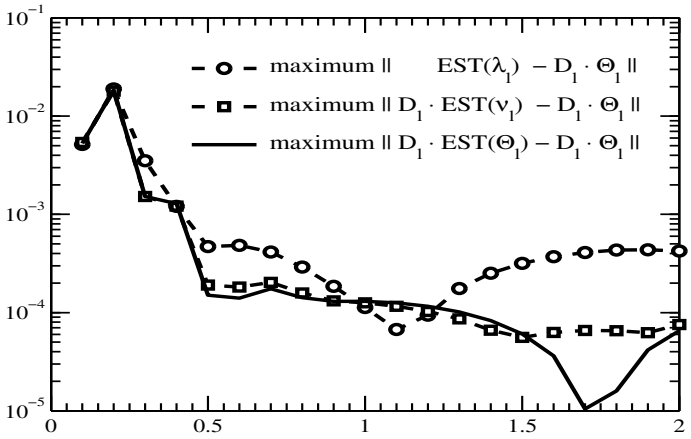


Figure 7.3: Regular index-1 DAE: The maximum norm of the difference between the exact local error Θ_ℓ and the local error estimates $\hat{\lambda}_\ell$, $D_\ell \hat{\nu}_\ell$, and $D_\ell \hat{\Theta}_\ell$ produced by the BDF_2 with constant stepsize $h = 0.1$ on the interval $[0, 2]$. $\text{EST}(\lambda_1)$, $\text{D}_1 \cdot \text{EST}(\nu_1)$, and $\text{D}_1 \cdot \text{EST}(\Theta_1)$ represent the error estimations $\hat{\lambda}_\ell$, $D_\ell \hat{\nu}_\ell$, and $D_\ell \hat{\Theta}_\ell$, respectively.

Example 7.8. Index-2 DAE with harmless critical points

Consider the DAE (3.16), discussed in Section 3.2.2, with

$$A(t) = \begin{bmatrix} \alpha(t) & 0 & 0 \\ 0 & \beta(t) & 0 \\ 0 & 0 & \gamma(t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, D = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

where $\alpha(t) = 1$, $\beta(t) = t^2 + 1$ and $\gamma(t) = \sin(4t)$. Since the zeros of $\gamma(t)$ define harmless critical points (cf. Example 3.16), this DAE has harmless critical points at $t = \pm\frac{n\pi}{4}$, $n = 0, 1, \ldots$. We set

$$\begin{aligned} x_1(t) &= \cos(4t), & x_2(t) &= -\sin(4t)\cos(4t), \\ x_3(t) &= 4(t^2+1)\sin^2(4t), & x_4(t) &= 4t^2\cos(8t), \quad x_5(t) = -8t^2\cos^2(4t), \end{aligned}$$

and then compute q. Consequently, the system of DAE reads

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & t^2+1 & 0 \\ 0 & 0 & \sin(4t) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_3(t) \\ x_2(t) \\ x_1(t) \end{bmatrix}' + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} x(t) =$$

$$\begin{bmatrix} (32t^2+31)\sin(4t)\cos(4t) + 8t\sin^2(4t) \\ -2(t^2+1) + 2(t^2-1)\cos(8t) \\ -2t^2(1+3\cos(8t)) \\ -4t^2 - \frac{1}{2}\sin(8t) \\ \cos(4t) + 4(t^2+1)\sin^2(4t) \end{bmatrix}. \tag{7.30}$$

We applied the BDF$_2$ on the interval $[0, 2]$ using `rtol` $= 10^{-j}$, $j = 1, \ldots, 8$ and `atol` $=$ `rtol` to the problem (7.30). The initial stepsize was chosen as $h_1 = 2 \cdot 10^{-3} \cdot$(`rtol` + `atol`). Figure 7.4 exhibited the achieved accuracy at the end of the integration interval (left) and the number of steps (right) with respect to the used tolerances. The comparison of the local error estimates $D_\ell\hat{\Theta}_\ell$ and $\hat{\Theta}_\ell$ specified in (7.16) has been illustrated for `rtol` $= 10^{-j}$, $j = 1, \ldots, 5$ in Figure 7.5. The accuracy is measured as the minimum number of significant correct digits in the numerical solution at the end of the integration interval, i.e

$$\text{accuracy} = \min(-\log_{10}(\| \text{ relative error at the end of the integration interval } \|_\infty)).$$

We observe that the computation by means of the error estimate $D_\ell\hat{\Theta}_\ell$ provides more accurate results than the one using the error estimate $\hat{\lambda}_\ell$ but takes more steps. The error estimates $D_\ell\hat{\nu}_\ell$ and $D_\ell\hat{\Theta}_\ell$ perform similarly for all tolerances. Furthermore, the estimator $D_\ell\hat{\Theta}_\ell$ is significantly better than the estimate $\hat{\Theta}_\ell$.

In the next section we investigate how the error estimates described above can be applied to the implicit Runge-Kutta method

$$x_\ell = X_{\ell s}, \tag{7.31a}$$

$$A_{\ell i}[DX]'_{\ell i} + B_{\ell i}X_{\ell i} = q_{\ell i}, \qquad i = 1, \ldots, s, \tag{7.31b}$$

where the internal derivatives $[DX]'_{\ell i}$ are defined by

$$[DX]'_{\ell i} = \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}(D_{\ell j}X_{\ell j} - D_{\ell-1}x_{\ell-1}), \qquad i = 1, \ldots, s, \tag{7.32}$$

with $A_{\ell i} := A(t_{\ell i})$, $D_{\ell i} := D(t_{\ell i})$, $B_{\ell i} := B(t_{\ell i})$, $q_{\ell i} := q(t_{\ell i})$, $i = 1, \ldots, s$.

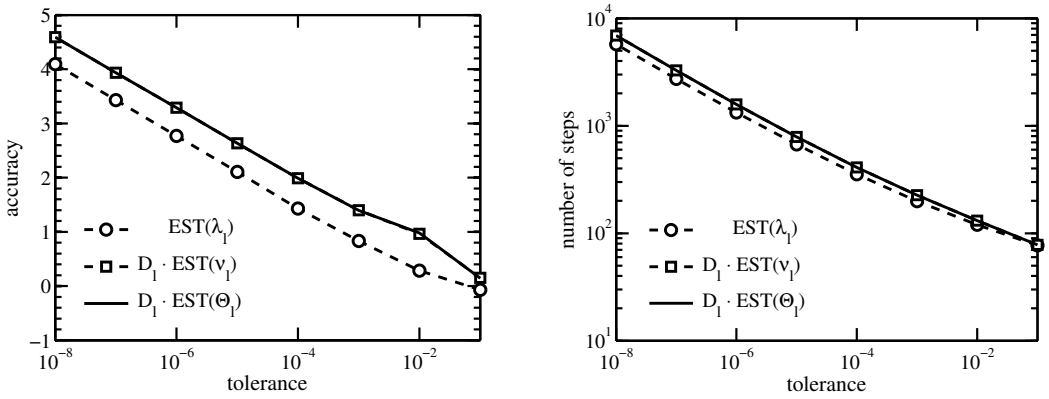


Figure 7.4: Index-2 DAE with harmless critical points: The accuracy (left) and the number of steps (right) provided by the BDF_2 using $\hat{\lambda}_\ell$ in (7.15), $D_\ell\hat{\nu}_\ell$ in (7.20), and $D_\ell\hat{\Theta}_\ell$ in (7.24), respectively, vs. the used tolerances. Lines with $\mathrm{EST}(\lambda_l)$, $\mathrm{D_l}\cdot\ \mathrm{EST}(\nu_l)$, and $\mathrm{D_l}\cdot\ \mathrm{EST}(\Theta_l)$ represent the error estimations $\hat{\lambda}_\ell$, $D_\ell\hat{\nu}_\ell$, and $D_\ell\hat{\Theta}_\ell$, respectively.

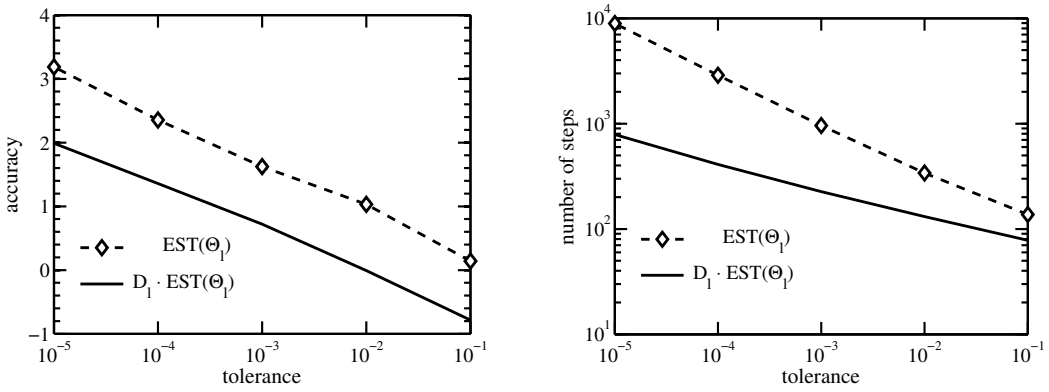


Figure 7.5: Index-2 DAE with harmless critical points: The accuracy (left) and the number of steps (right) provided by the BDF_2 using $\hat{\Theta}_\ell$ in (7.16) and $D_\ell\hat{\Theta}_\ell$ in (7.24) vs. the used tolerances. Lines with $\mathrm{EST}(\Theta_l)$, $\mathrm{D_l}\cdot\ \mathrm{EST}(\Theta_l)$ represent the estimates $\hat{\Theta}_\ell$ and $D_\ell\hat{\Theta}_\ell$, respectively.

7.2 Error estimation and stepsize prediction for IRK methods

As in the previous section, we specify the local error and the local truncation error for the IRK methods (7.31). The local error for the IRK method (7.31) is defined as

$$\Theta_{\ell i} \;:=\; X^*_{\ell i} - x_*(t_{\ell i}), \qquad i = 1, \ldots, s, \tag{7.33}$$

where $x_*(t)$ is the exact solution of the DAE (6.1) and the stage approximations $X^*_{\ell 1}$, …, $X^*_{\ell s}$, are obtained by solving the discretized equation (7.31b) from the value $x_*(t_{\ell-1})$ replacing $x_{\ell-1}$.

Moreover, the local truncation error of the IRK method (7.31) is denoted by

$$\lambda_{\ell i} \;:=\; h\left((Dx_*)'(t_{\ell i}) - [(Dx_*)]'_{\ell i}\right), \qquad i = 1, \ldots, s, \tag{7.34}$$

where $[(Dx_*)]'_{\ell i} := \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}\left(D_{\ell j}x_*(t_{\ell j}) - D_{\ell-1}x_*(t_{\ell-1})\right)$.

The relation between the errors $\Theta_{\ell i}$ and $\lambda_{\ell i}$ is represented in the following statement.

Lemma 7.9. *Let the stepsize* $\boldsymbol{h}$ *be sufficiently small to guarantee that the matrix*

$$\left(\frac{1}{h}(\mathcal{A}^{-1} \otimes I)A_\ell D_\ell + B_\ell\right) \tag{7.35}$$

is nonsingular. Then, the errors $\Theta_{\ell i}$ *and* $\lambda_{\ell i}$ *satisfy the relation*

$$\Theta_\ell \;=\; \left(\frac{1}{h}(\mathcal{A}^{-1} \otimes I)A_\ell D_\ell + B_\ell\right)^{-1} \cdot \frac{1}{h}\,\mathcal{D}_A\,\lambda_\ell, \tag{7.36}$$

where $\mathcal{A}^{-1} = (\tilde{\alpha}_{ij})_{i,j=1}^{s}$, $\mathcal{D}_A := \operatorname{diag}(A_{\ell 1}, \ldots, A_{\ell s})$, *and* $\Theta_\ell := (\Theta_{\ell 1}, \ldots, \Theta_{\ell s})^T$, $\lambda_\ell := (\lambda_{\ell 1}, \ldots, \lambda_{\ell s})^T$, *and similar for* A_ℓ, D_ℓ, *and* B_ℓ.

Proof : From the definition (7.34) we may write $(Dx_*)'(t_{\ell i})$ as

$$(Dx_*)'(t_{\ell i}) \;=\; \frac{1}{h}\lambda_{\ell i} + \frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}\left(D_{\ell j}x_*(t_{\ell j}) - D_{\ell-1}x_*(t_{\ell-1})\right), \quad i = 1, \ldots, s. \tag{7.37}$$

Since the exact solution $x_*(t)$ satisfies

$$A_{\ell i}(Dx_*)'(t_{\ell i}) + B_{\ell i}x_*(t_{\ell i}) - q_{\ell i} \;=\; 0, \quad i = 1, \ldots, s, \tag{7.38}$$

inserting (7.37) into (7.38) yields

$$A_{\ell i}\frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}\left(D_{\ell j}x_*(t_{\ell j}) - D_{\ell-1}x_*(t_{\ell-1})\right) + B_{\ell i}x_*(t_{\ell i}) - q_{\ell i} \;=\; -\frac{1}{h}A_{\ell i}\lambda_{\ell i}. \tag{7.39}$$

Since $X^*_{\ell i}$, $i = 1, \ldots, s$, satisfy the relation

$$A_{\ell i}\frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}\left(D_{\ell j}X^*_{\ell j} - D_{\ell-1}x_*(t_{\ell-1})\right) + B_{\ell i}X^*_{\ell i} - q_{\ell i} \;=\; 0, \quad i = 1, \ldots, s, \tag{7.40}$$

then subtracting (7.40) by (7.39) we obtain

$$\begin{aligned} A_{\ell i}\frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}D_{\ell j}\left(X^*_{\ell j} - x_*(t_{\ell j})\right) + B_{\ell i}\left(X^*_{\ell i} - x_*(t_{\ell i})\right) \;&=\; \frac{1}{h}A_{\ell i}\lambda_{\ell i}, \\ A_{\ell i}\frac{1}{h}\sum_{j=1}^{s} \tilde{\alpha}_{ij}D_{\ell j}\Theta_{\ell j} + B_{\ell i}\Theta_{\ell i} \;&=\; \frac{1}{h}A_{\ell i}\lambda_{\ell i}. \end{aligned} \tag{7.41}$$

Indeed, (7.41) can be written as

$$\begin{bmatrix} \frac{\tilde{\alpha}_{11}}{h}A_{\ell 1}D_{\ell 1}+B_{\ell 1} & \frac{\tilde{\alpha}_{12}}{h}A_{\ell 1}D_{\ell 2} & \cdots & \frac{\tilde{\alpha}_{1s}}{h}A_{\ell 1}D_{\ell s} \\ \frac{\tilde{\alpha}_{21}}{h}A_{\ell 2}D_{\ell 1} & \frac{\tilde{\alpha}_{22}}{h}A_{\ell 2}D_{\ell 2}+B_{\ell 2} & \cdots & \frac{\tilde{\alpha}_{2s}}{h}A_{\ell 2}D_{\ell s} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\tilde{\alpha}_{s1}}{h}A_{\ell s}D_{\ell 1} & \frac{\tilde{\alpha}_{s2}}{h}A_{\ell s}D_{\ell 2} & \cdots & \frac{\tilde{\alpha}_{ss}}{h}A_{\ell s}D_{\ell s}+B_{\ell s} \end{bmatrix} \begin{bmatrix} \Theta_{\ell 1} \\ \Theta_{\ell 2} \\ \vdots \\ \Theta_{\ell s} \end{bmatrix}$$
$$= \frac{1}{h}\begin{bmatrix} A_{\ell 1}\lambda_{\ell 1} \\ A_{\ell 2}\lambda_{\ell 2} \\ \vdots \\ A_{\ell s}\lambda_{\ell s} \end{bmatrix}. \tag{7.42}$$

With the notations

$$\mathcal{D}_A := \mathrm{diag}(A_{\ell 1},\ldots,A_{\ell s}), \quad \Theta_\ell := (\Theta_{\ell 1},\ldots,\Theta_{\ell s})^T, \quad \lambda_\ell := (\lambda_{\ell 1},\ldots,\lambda_{\ell s})^T,$$

and similar for A_ℓ, D_ℓ, and B_ℓ, one may write (7.42) in a compact form to obtain

$$\left(\frac{1}{h}(\mathcal{A}^{-1}\otimes I)A_\ell D_\ell + B_\ell\right)\Theta_\ell \;=\; \frac{1}{h}\,\mathcal{D}_A\,\lambda_\ell,$$

and therefore the statement is verified. □

Remark 7.10. *As in Chapter 6, if we define the local error as the defect obtained when the exact solution is inserted into the numerical scheme, i.e.*

$$\tau_{\ell i} \;:=\; A_{\ell i}\frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}\,(D_{\ell j}x_*(t_{\ell j}) - D_{\ell-1}x_*(t_{\ell-1})) + B_{\ell i}x_*(t_{\ell i}) - q_{\ell i}, \quad i=1,\ldots,s,$$

then the local error $\tau_{\ell i}$ relates to $\lambda_{\ell i}$ as follow:

$$\begin{aligned}\tau_{\ell i} \;&=\; A_{\ell i}\frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}\,(D_{\ell j}x_*(t_{\ell j}) - D_{\ell-1}x_*(t_{\ell-1})) - A_{\ell i}(Dx_*)'(t_{\ell i}), \quad i=1,\ldots,s \\ &=\; -\frac{1}{h}A_{\ell i}\lambda_{\ell i}, \quad i=1,\ldots,s.\end{aligned}$$

Therefore, the error $\Theta_{\ell i}$ and defect $\tau_{\ell i}$ satisfy the relation

$$A_{\ell i}\frac{1}{h}\sum_{j=1}^{s}\tilde{\alpha}_{ij}D_{\ell j}\Theta_{\ell j} + B_{\ell i}\Theta_{\ell i} \;=\; -\tau_{\ell i}.$$

As above (cf. Equation (7.42)), denoting $\tau_\ell := (\tau_{\ell 1},\ldots,\tau_{\ell s})^T$ we obtain

$$\Theta_\ell \;=\; -\left(\frac{1}{h}(\mathcal{A}^{-1}\otimes I)A_\ell D_\ell + B_\ell\right)^{-1}\tau_\ell,$$

if the matrix $\left(\frac{1}{h}(\mathcal{A}^{-1}\otimes I)A_\ell D_\ell + B_\ell\right)$ is nonsingular.

Summary

Differential-algebraic equations (DAEs) describe dynamical processes that are restricted by some constraints and arise in numerous fields of applications. The numerical treatment of DAEs requires knowledge about their structure. In contrast to explicit regular ordinary differential equations, DAEs involve not only integration problems but also differentiation problems. Differentiation problems are ill-posed in the sense that small perturbations in the initial data may cause arbitrarily large defects in the solution data. Hence, appropriate numerical computations are required.

We investigated numerical integration methods for general linear index-2 DAEs

$$A(t)(D(t)x(t))' + B(t)x(t) \;=\; q(t), \tag{7.43}$$

having harmless critical points. Under the application of quasi-admissible projector functions instead of the admissible ones, besides DAEs which have almost everywhere the same characteristic values, DAEs with index changes can now be discussed for the first time and harmless critical points of a linear DAE (7.43) can be characterized.

The special implicit Runge-Kutta method (see page 69) and the BDF scheme are proved to be only weakly stable. Further, by using the decoupling procedure convergence for general linear index-2 DAEs (7.43) with harmless critical points is proven. The convergence proof stated that the implicit Runge-Kutta and BDF methods achieve the same order of convergence for this class of DAEs as they do for ordinary differential equations. The weak instability affects only certain components and is not accumulated. For quasi-regular index-1 DAEs no influence of the weak instability is visible.

Using the decoupling procedure, an appropriate local error estimate for the BDF method applied to linear index-2 DAE (7.43) can be derived. The problems which are due to the instable derivative-free part can be avoided in the error control if we scale the error estimate by the coefficient matrix D of the DAE. The numerical results confirm that the method performs very good when the error control is based only on the differential components.

Bibliography

[1] K. Balla and V. H. Linh. Adjoint pairs of differential-algebraic equations and hamiltonian ststems. *Appl. Numer. Math.*, 53:131–148, 2005.

[2] K. Balla and R. März. A unified approach to linear differential algebraic equations and their adjoint equations. *J. Anal. Appl.*, 21:783–802, 2002.

[3] K. Balla, G. A. Kurina, and R. März. Index criteria for differential algebraic equations arising form linear-quadratic optimal control problems. *J. Dyn. Control Syst.*, 12:289–311, 2006.

[4] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer-Verlag, 2003.

[5] M. Berzins and R. M. Furzeland. A user's manual for SPRINT - a versatile software package for solving systems of algebraic, ordinary and partial differential equations: part 1 - algebraic and ordinary differentail equations. *Thornton Research Centre, Shell Research Ltd.*, 1985. TNER.85.058.

[6] R. K. Brayton, F. G. Gustavson, and G. D. Hachtel. A new efficient algorithm for solving differential-algebraic systems using implicit backward differentiation formulas. *Proc. IEEE*, 60:98–108, 1972.

[7] K. E. Brenan. *Stability and Convergence of Difference Approximations for Higher Index Differential-Algebraic Systems with Applications in Trajectory Control*. PhD thesis, University of California, Los Angeles, 1983.

[8] K. E. Brenan and L. R. Engquist. Backward differentiation approximations of nonlinear differential/algebraic equations. *Math. Comp.*, 51:659–676, 1988.

[9] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics (SIAM), North-Holland, New York, 1989.

[10] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*, volume 14 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), 1996. ISBN 90-89871-353-6.

[11] R. L. Brown and C. W. Gear. Documentation for DFASUB - a program for the solution of simultaneous implicit differential and non-linear equations. Technical Report UIUCDCS-R-73-575, University of Illinois at Urbana-Champaign, 1973.

[12] J. C. Butcher. *The numerical analysis of ordinary differential equations, Runge-Kutta and general linear methods*. Wiley, Chichester and New York, 1987.

[13] J. C. Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons Ltd., Chichester, 2003. ISBN 0-471-96758-0.

[14] J. C. Butcher and A. D. Heard. Stability of numerical methods for ordinary diferential equations. *Numer. Algorithms*, 31(1-4):59–73, 2002.

[15] C. W. Cryer. On the instability of high order backward-difference multistep methods. *BIT*, 12:17–25, 1972.

[16] C. F. Curtiss and J. O. Hirschfelder. Integration of stiff equations. *Proc. Nat. Acad. Sci.*, 38:235–243, 1952.

[17] G. Denk. Die numerische Integration von Algebro-Differentialgleichungen bei der Simulation elektrischer Schaltkreise mit SPICE2. Technical Report TUM-M8809, Technische Universität München, 1988.

[18] H. Döring. Traktabilitätsindex und Eigenschaften von matrixwertigen Riccati-Typ Algebrodifferentialgleichungen. Master's thesis, Humboldt-Universität zu Berlin, Institut für Mathematik, Berlin, 2004.

[19] B. L. Ehle. On Pade approximations to the exponential function and A-stable methods for the numerical solution of initial value problems. Technical report, Dept. AACS, University of Waterloo, Ontario, Canada, 1969. Research Report CSRR 2010.

[20] J. Flügel. Lösung von Algebro-Differentialgleichungen mit properem Haupterm durch die BDF. Master's thesis, Humboldt-Universität zu Berlin, Institut für Mathematik, Berlin, 2002.

[21] F. R. Gantmacher. *The Theory of Matrices*. Chelsea Pub. Co., New York, 1959.

[22] C. W. Gear. The simultaneous numerical solution of differential-algebraic equations. *IEEE Trans. Circuit and Theory*, CT-18(1):89–95, 1971.

[23] C. W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice Hall, Englewood Cliffs, NJ, 1971.

[24] C. W. Gear and L. R. Petzold. Differential/algebraic systems and matrix pencils. In B. Kagstrom and A. Ruhe, editors, *Matrix Pencils*, volume 973 of *Lecture Notes in Mathematics*, pages 75–89. Springer Verlag, 1983.

[25] C. W. Gear and L. R. Petzold. ODE methods for the solution of differential/algebraic systems. *SIAM J. Numer. Anal.*, 21:716–728, 1984.

[26] C. W. Gear, G. K. Gupta, and B. Leimkuhler. Automatic integration of Euler Lagrange equations with constraints. *Comp. Appl. Math.*, 12 & 13:77–90, 1985.

[27] E. Griepentrog and R. März. *Differential-algebraic equations and their numerical treatment*. Number 88 in Teubner Texte zur Mathematik. Teubner, Leipzig, 1986.

[28] E. Griepentrog and R. März. Basic properties of some differential-algebraic equations. *Z. Anal. Anwendungen*, 8:25–40, 1989.

[29] R. D. Grigorieff. *Numerik gewöhnlicher Differentialgleichungen 2*. Teubner Studienbücher Mathematik. B. G. Teubner, Stuttgart, 1977.

[30] K. Gustafsson. Control-theoretic techniques for stepsize selection in implicit Runge-Kutta methods. *ACM Trans. Math. Software*, 20:496–517, 1994.

[31] K. Gustafsson, M. Lundh, and G. Söderlind. A PI stepsize control for the numerical solution of ordinary differential equations. *BIT Numerical Mathematics*, 28:270 – 287, 1988.

[32] E. Hairer. Radau5: Implicit Runge-Kutta method of order 5 (Radau IIA) for semi-implicit DAEs. `http://www.unige.ch/~hairer/software.html`.

[33] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer, Berlin Heidelberg New York Tokyo, 1991.

[34] E. Hairer and G. Wanner. Stiff differential equations solved by Radau methods. *J. Comput. Appl. Math.*, 111:93–111, 1999.

[35] E. Hairer, C. Lubich, and R. Rocher. *The numerical solution of differential-algebraic systems by Runge-Kutta methods*, volume 1409 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1989. ISBN 3-540-51860-6.

[36] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 2. rev. edition of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1993.

[37] G. Hall and J. M. Watt. *Modern Numerical Methods for Ordinary Differential Equations*. Clarendon Press, Oxford, 1976.

[38] M. Hanke. A new implementation of a BDF method within the method of lines. *Comput. Meth. Sci. Engg.*, To appear 2003.

[39] K. Heun. Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen. *Z. für Math. u. Phys.*, 45:23–38, 1900.

[40] I. Higueras and R. März. Formulating differential algebraic equations properly. Technical Report 00-20, Humboldt-Universität zu Berlin, Institut für Mathematik, 2000.

[41] I. Higueras and R. März. Differential algebraic equations with properly stated leading terms. *Comp. Math. Appl.*, 48:215–235, 2004.

[42] I. Higueras, R. März, and C. Tischendorf. Stability preserving integration of index-1 DAEs. *Appl. Num. Math.*, 45:175–200, 2003.

[43] I. Higueras, R. März, and C. Tischendorf. Stability preserving integration of index-2 DAEs. *Appl. Num. Math.*, 45:201–229, 2003.

[44] A. C. Hindmarsh. ISODE and LSODI, two new initial value ordinary differential equation solvers. *ACM-SIGNUM Newsletter*, 15:10–11, 1980.

[45] R. Horn and C. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[46] K. R. Jackson and R. Sacks-Davis. An alternative implementation of variable step-size multistep formulas for stiff ODEs. *ACM Trans. Math. Software*, 6: 295–318, 1980.

[47] O. Koch, R. März, D. Praetorius, and E. Weinmüller. Collocation methods for index 1 DAEs with a singularity of the first kind. *Math. Comp.*, 79(269): 281–304, 2010.

[48] L. Kronecker. Algebraische Reduction der Schaaren bilinearer Formen. *Sitzungsberichte Akad. Wiss. Berlin*, pages 1225–1237, 1890.

[49] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations. Analysis and Numerical Solution*. EMS Publishing House, Zürich, Switzerland, 2006.

[50] G. A. Kurina and R. März. On linear-quadratic optimal control problems for time-varying descriptor systems. *SIAM J. Control Optimization*, 42:2062–2077, 2004.

[51] W. Kutta. Beitrag zur näherungsweisen Integration totaler Differentialgleichungen. *Z. Math. Phys.*, 46:435–453, 1901.

[52] A. Kvaerno. Runge-Kutta methods applied to fully implicit differentail-algebraic equations of index 1. *Math. Comp.*, 54(190):583–625, 1990. ISSN 0025-5718.

[53] J. D. Lambert. *Computational Methods in Ordinary Differential Equations*. Wiley, New York, 1973.

[54] R. Lamour, R. März, and C. Tischendorf. Projector Based DAE Analysis. book in preparation.

[55] P. Lötstedt and L. R. Petzold. Numerical solution of nonlinear differential equations with algebraic constraints I: Convergence results for backward differentiation formulas. *Math. Comp.*, 46(174):491–516, 1986.

[56] E. I. Macana. *Numerische Approximation von Algebro-Differentialgleichungen mit Index 2 mittels impliziter Runge-Kutta-Verfahren.* PhD thesis, Humboldt-Universität zu Berlin, Institut für Mathematik, 1993.

[57] R. März. A matrix chain for analyzing differential algebraic equations. Technical Report 162, Humboldt-Universität zu Berlin, Institut für Mathematik, 1987.

[58] R. März. Index-2 differential-algebraic equations. *Results in Mathematics*, 15: 149–171, 1989.

[59] R. März. Some new results concerning index-3 differential-algebraic equations. *J. Mathe. Analysis and Applications*, 140(1):177–199, 1989.

[60] R. März. Numerical methods for differential algebraic equations. *Acta Numerica*, pages 141–198, 1992.

[61] R. März. Canonical projectors for linear differential algebraic equations. *Comp. Math. Appl.*, 31:121–135, 1996.

[62] R. März. Adjoint equations of differential-algebraic systems and optimal control problems. In *Proc. Inst. Math.*, volume 7, pages 88–97. NAS Belarus, 2001.

[63] R. März. Differential algebraic equations anew. *Appl. Numer. Math.*, 42: 315–335, 2002.

[64] R. März. The index of linear differential algebraic equations with properly stated leading terms. *Results in Mathematics*, 42:308–338, 2002.

[65] R. März. Solvability of linear differential algebraic equations with properly stated leading terms. *Results in Mathematics*, 45:88–105, 2004.

[66] R. März. Fine decouplings of regular differential algebraic equations. *Results in Mathematics*, 46:57–72, 2004.

[67] R. März. Projector Based DAE Analysis. Technical report, Mathematisches Forschungsinstitut Oberwolfach, 2006.

[68] R. März and R. Riaza. Linear differential-algebraic equations with properly stated leading term: Regular points. *J. Math. Anal. Appl.*, 323:1279–1299, 2006.

[69] R. März and R. Riaza. Linear differential-algebraic equations with properly stated leading term: A-critical points. *Math. Comput. Model. Dyn. Syst.*, 13: 291–314, 2007.

[70] R. März and R. Riaza. Linear differential-algebraic equations with properly stated leading term: B-critical points. Technical Report 07-9, Humboldt-Universität zu Berlin, Institut für Mathematik, 2007.

[71] R. März and R. Riaza. Linear differential-algebraic equations with properly stated-leading term: B-critical points. *Dynamical Systems*, 23:505–522, 2008.

[72] L. R. Petzold. A description of DASSL: A differential/algebraic system solver. In R. S. Stepleman et al., editor, *Scientific Computing*, pages 65–68. Elsevier, North-Holland, Amsterdam, 1983.

[73] L. R. Petzold. Order results for implicit Runge-Kutta methods applied to differential/algebraic systems. *SIAM J. Numer. Anal.*, 23:837–852, 1986.

[74] A. Prothero and A. Robinson. On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Math. Comp.*, 28:145–162, 1974.

[75] P. J. Rabier and W. C. Rheinboldt. Theoretical and numerical analysis of differential-algebraic equations. In *Techniques of Scientific Computing (Part 4)*, volume 8 of *Handbook of Numerical Analysis*, pages 183 – 540. Elsevier, 2002.

[76] R. Riaza. *Differential-Algebraic Systems: Analytical Aspects and Circuit Applications*. World Scientific, Singapore, 2008.

[77] R. Riaza and R. März. Linear index-1 DAEs: regular and singular problems. *Acta Applicandae Mathematicae*, 84:29–53, 2004.

[78] M. Roche. Rosenbrock methods for differential algebraic equations. *Numer. Math.*, 52:45–63, 1988.

[79] T. Rübner-Peterson. An efficient algorithm using backward time-scaled differences for solving stiff differential algebraic systems. Technical report, Institute of Circuit Theory and Telecommunication, Technical University of Denmark, 2800 Lyngby, 1973.

[80] C. Runge. Über die numerische Auflösung von Differentialgleichungen. *Math. Ann.*, 46:167–178, 1895.

[81] S. Schulz. Four Lectures on Differential-Algebraic Equations. Technical Report 497, University of Auckland, Department of Mathematics, 2003.

[82] J. Sieber. Fehlerkontrolle und Schrittweitensteuerung bei der numerischen Integration von Algebro-Differentialgleichungen mit der BDF. Master's thesis, Humboldt-Universität zu Berlin, Institut für Mathematik, 1997.

[83] J. Sieber. Local error control of general index-1 and index-2 differential-algebraic equations. Technical Report 97-21, Humboldt-Universität zu Berlin, Institut für Mathematik, 1997.

[84] R. F. Sincover, A. M. Erisman, E. L. Yip, and M. A. Epton. Analysis of descriptor systems using numerical algorithms. *IEEE Trans. Automat. Control*, AC-26:139–147, 1981.

[85] A. Sjö. *Analysis of computational algorithms for linear multistep methods*. PhD thesis, Mathematical Sciences, Numerical Analysis, Centre for Mathematical Sciences, Lund University, 1999.

[86] G. Söderlind. Digital filters in adaptive time-stepping. *ACM Trans. Math. Software*, 29:1–26, 2003.

[87] G. Söderlind. Time-step algorithms: Adaptivity, control and signal processing. *Appl. Num. Math.*, 56:488–502, 2006.

[88] J. Stoer. *Einführung in die Numerische Mathematik I.* Springer-Verlag, 4th edition, 1983.

[89] C. Tischendorf. Die BDF für nichtlineare Algebro-Differentialgleichungen vom Index 2. Master's thesis, Humboldt-Universität zu Berlin, Institut für Mathematik, 1992.

[90] C. Tischendorf. Feasibility and stability behaviour of the BDF applied to index-2 differential algebraic equations. *ZAMM*, 75(12):927–946, 1995.

[91] C. Tischendorf. *Solution of index-2 differential algebraic equations and its application in circuit simulation*. PhD thesis, Humboldt-Universität zu Berlin, Institut für Mathematik, 1996.

[92] G. Zielke. Motivation und Darstellung von verallgemeinerten Matrixinversen. *Beiträge zur Numerischen Mathematik*, 7:177–218, 1979.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Dissertation selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Berlin, den 12. Januar 2011